LOGICS OF BELIEF

by

ELIZABETH VILJOEN

submitted in part fulfilment of the requirements
for the degree of

MASTER OF SCIENCE

in the subject

COMPUTER SCIENCE

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF W A LABUSCHAGNE

APRIL 1997

Abstract

The inadequacy of the usual possible world semantics of modal languages when the meaning of 'belief' is attached to the modal operator is discussed. Three other approaches are then investigated. In the case of Moore's autoepistemic logic it becomes possible to compare an agent's beliefs to 'reality', which cannot be done directly in the possible world semantics. Levesque's semantics makes explicit in the object language the notion of 'this is all the information the agent has', which plays an important role in nonmonotonic reasoning. Both of these approaches deal with ideal reasoners. The third approach, Konolige's deduction model, is based on a semantics capable of describing the beliefs of one or more resource-bounded agents. Finally, the AGM postulates for belief revision are discussed.

Key terms Logic of belief; Epistemic logic; Possible worlds; Autoepistemic logic; Agent; Ideal reasoner; Only knowing; Resource-bounded reasoner; Deduction model; Belief revision; AGM postulates

"Om te glo, is om seker te wees van die dinge wat ons hoop,
om oortuig te wees van die dinge wat ons nie sien nie."
Hebrews 11:1

Acknowledgements

I would like to express my sincere thanks and appreciation to each of the following:

- prof. W.A. Labuschagne of the Department of Computer Science and Information Systems at Unisa for his unfailing enthusiasm, patience and expert guidance,
- prof. J. Heidema of the Department of Mathematics, Applied Mathematics and Astronomy at Unisa for his valuable comments,
- the Department of Computer Science and Information Systems at Unisa for granting me study leave and allowing me the use of their facilities,
- my colleagues at the department for their invaluable assistance and positive attitude,
- René, Henning, Henda, Maretha and Stefanie for their love, support, interest and understanding when 'their' time was used for studies and
- our Heavenly Father without Whom nothing would have been possible – 'unless the Lord builds the house, its builders labour in vain; unless the Lord watches over the city, the watchmen stand guard in vain'.

Biffie Viljoen
April 1997

Contents

Chapter 1

Classical Background

Why is the sentence
'p but I do not believe that p'
absurd to utter?
Moore's Problem as stated by Hintikka (1962)

What motivates the study of logics of belief?

The ultimate goal of the logician who devises a logic of belief is to give a formal description of the kind of set of beliefs that a rational agent might have and, with the help of this explication, to establish criteria for distinguishing between good and bad arguments involving the notion of belief. The explication must take into account the manner in which beliefs are adopted, supported and revised by real agents, but the logician's purpose is to develop a normative theory, not to give a psychological analysis.

1.1 Propositional Logics

Typically, logics of knowledge and belief are formulated in terms of modal languages. For the sake of simplicity, we will concentrate on propositional languages. Such languages are extensions of the languages of classical propositional logic. Accordingly we first discuss the syntax, model theory and proof theory of classical propositional languages.

1.1.1 Syntax

The first requirement for any formal language is an alphabet $\mathcal{S}$ of symbols. For a propositional language these symbols include atoms, connectives and parentheses for punctuation. The set of *atoms* $\mathcal{P}$ may have any non-zero cardinality although it is customary to choose a countable set. For illustrative purposes we will use $\mathcal{P} = \{p, q\}$. Given $\mathcal{P}$, the *alphabet* of the propositional language is the set

$\mathcal{S} = \mathcal{P} \cup \{\neg, \vee, \wedge, \rightarrow, \leftrightarrow, (,)\}$.

Let $\mathcal{S}^*$ be the set of all possible strings over $\mathcal{S}$. The *well-formed formulae* (wfs) or *sentences* of the propositional language are the members of the set $\mathcal{F}$ defined as follows:

$\mathcal{F}$ is the smallest subset of $\mathcal{S}^*$ such that

- $\mathcal{P} \subseteq \mathcal{F}$,
- if $A \in \mathcal{F}$, then $\neg A \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \vee B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \wedge B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \rightarrow B) \in \mathcal{F}$ and
- if $A, B \in \mathcal{F}$, then $(A \leftrightarrow B) \in \mathcal{F}$.

For easier readability we may omit parentheses if no ambiguity can arise.

Examples of well-formed formulae in the language where $\mathcal{P} = \{p, q\}$ are
$\neg\neg q$
$(p \rightarrow (p \vee q))$ or, omitting some parentheses, $p \rightarrow (p \vee q)$
$\neg(p \wedge \neg p)$
$((q \leftrightarrow p) \rightarrow (p \leftrightarrow q))$.

1.1.2 Model Theory

We have looked at the syntax of classical propositional languages. Now we take a look at the semantics of such languages. Our basic intuition is that when a sentence accurately describes a situation (or world or state) of a system, this sentence is true with respect to that situation or world or state. Whereas the semantics of predicate languages (to be discussed later) will make explicit the worlds, we will be content here to represent worlds by indicating which elementary facts (i.e. atoms) hold or do not hold in each of these worlds.

The elementary facts holding in a world are indicated by an *assignment* of truth values to atoms, namely a function f with domain the set of atoms, $\mathcal{P}$, and as range the set {T,F} of truth values. The atoms p such that $f(p)$=T are then the elementary facts that hold in the world that is implicitly under discussion.

An assignment of truth values to atoms can be extended in many ways to a function with domain the set of all wfs of the language. There is, however, only one way to extend an assignment in such a manner that our intuitions with regard to the connectives are respected. The *valuation* v_f extending the assignment f is recursively defined as follows:

$$
\begin{aligned}
v_f(p) &= f(p) \text{ for every } p \in \mathcal{P} \\
v_f(\neg B) &= \begin{cases} \text{T} & \text{if } v_f(B) = \text{F} \\ \text{F} & \text{if } v_f(B) = \text{T} \end{cases} \\
v_f((B \vee C)) &= \begin{cases} \text{T} & \text{if } v_f(B) = \text{T or } v_f(C) = \text{T or both} \\ \text{F} & \text{if } v_f(B) = v_f(C) = \text{F} \end{cases} \\
v_f((B \wedge C)) &= \begin{cases} \text{T} & \text{if } v_f(B) = v_f(C) = \text{T} \\ \text{F} & \text{if } v_f(B) = \text{F or } v_f(C) = \text{F or both} \end{cases} \\
v_f((B \rightarrow C)) &= \begin{cases} \text{T} & \text{if } v_f(B) = \text{F or } v_f(C) = \text{T or both} \\ \text{F} & \text{if } v_f(B) = \text{T and } v_f(C) = \text{F} \end{cases} \\
v_f((B \leftrightarrow C)) &= \begin{cases} \text{T} & \text{if } v_f(B) = v_f(C) \\ \text{F} & \text{if } v_f(B) \neq v_f(C). \end{cases}
\end{aligned}
$$

If $v_f(A)$=T we say that A is 'true' in the world (represented by) f and that the specific assignment f *satisfies* the wf A. If $v_f(A)$ =F we say that A is 'false' in the world (represented by) f and that the assignment f fails to satisfy the wf A. The world (represented by) f is a *model* of a wf A if and only if A is true in f. This means that an assignment satisfies a wf A iff the world represented by the assignment is a model of A.

The world (represented by) f is a model of a set Γ of wfs if and only if every member of Γ is true in f. Similarly, an assignment f satisfies a set Γ of wfs iff it satisfies every wf of the set, i.e. an assignment satisfies a set Γ iff the world represented by the assignment is a model of the set Γ.

Let $\mathcal{V}$ be the set of all possible assignments, i.e. worlds. Let us further agree to call subsets of $\mathcal{V}$ *frames*. Then A is *valid in the frame* $\mathcal{W} \subseteq \mathcal{V}$ iff $v_f(A)$ =T for all worlds $f \in \mathcal{W}$.

Let us look at an example. In the language with only two atoms, p and q, we can have four possible worlds:

$f_1(p)$ =T and $f_1(q)$ =T
$f_2(p)$ =T and $f_2(q)$ =F
$f_3(p)$ =F and $f_3(q)$ =T
$f_4(p)$ =F and $f_4(q)$ =F.

In this case, then, $\mathcal{V} = \{f_1, f_2, f_3, f_4\}$. The wf $(p \vee \neg q)$ is true in worlds f_1, f_2 and f_4 or, equivalently, the wf $(p \vee \neg q)$ is valid in the frame $\{f_1, f_2, f_4\}$. This means that the worlds f_1, f_2 and f_4 are all models of the wf $(p \vee \neg q)$. The wf q, the wf $(p \rightarrow q)$ and the wf $(p \vee q)$ are all valid in the frame $\{f_1, f_3\}$, so the worlds f_1 and f_3 are models of the set $\{q, (p \rightarrow q), (p \vee q)\}$. The wf $(q \vee \neg q)$ is valid in the frame $\mathcal{V}$.

Every wf, and indeed every set of wfs, determines a particular frame, namely the set of worlds that are models of the specific wf or set of wfs.

A wf A *entails* a wf B , written as $A \models B$, iff B is valid in the frame determined by A, i.e. iff B is true in every model of A. The manner in which we assign truth values to wfs of the form $(A \rightarrow B)$ ensures that $A \models B$ iff the wf $(A \rightarrow B)$ is true in all worlds $w \in \mathcal{V}$, i.e. $(A \rightarrow B)$ is valid in the (whole) $\mathcal{V}$. More generally, a set Γ of wfs entails a wf A iff A is valid in the frame determined by Γ, i.e. A is true in every world in which each member of Γ is true. Unsurprisingly we write it as $\Gamma \models A$. This leads to the definition of the 'closure' of a set. The *semantic closure* $\mathrm{Cn}(\Gamma)$ of a set Γ is the set of all wfs that are true in all the models of Γ, i.e. the semantic closure of a set consists of all the wfs entailed by the set. So $\mathrm{Cn}(\Gamma)$ is the set of all wfs that are valid in the frame determined by Γ. The operator Cn is *monotonic*, i.e. given two sets Φ and Γ, if $\Phi \subseteq \Gamma$, then $\mathrm{Cn}(\Phi) \subseteq \mathrm{Cn}(\Gamma)$.

1.1.3 Proof Theory

For many purposes the entailment relation $\models$ is the basic relation of interest. To determine whether a pair (A, B) belongs to the relation (i.e. whether $A \models B$), we may do a model-theoretic analysis. Say, however, we want to automate the process. Then it becomes reasonable to investigate the existence of syntactic algorithms which would determine whether $A \models B$ by systematically rewriting the strings A and B instead of embarking on the complexities of a model-theoretic analysis. Before the advent of computer science, such syntactic algorithms were under-specified: the product to be delivered was specified, namely a 'proof', which was defined as a certain kind of sequence of wfs; the manner in which the proof was to be generated was left unspecified. In order to facilitate the comparison of

several approaches, we adhere to this tradition.

We have seen that $A \models B$ iff B is valid in the frame consisting of the models of A (and, more generally, $\Gamma \models B$ iff B is valid in the frame of all the models of Γ). Say we want to investigate validity in a frame which is determined by a wf A (or a set Γ of wfs). Then we refer to A as an *axiom* (or to Γ as a *set of axioms*).

Consider first the special case in which we wish to investigate validity in the frame $\mathcal{V}$ of all possible worlds. Are there axioms which may be regarded as determining this frame? Well, the obvious answer is that the set Γ consisting of all wfs valid in $\mathcal{V}$ constitutes such a set of axioms. But this renders the question 'Is it the case that $\Gamma \models B$?' trivial. It would be far more useful if there were a relatively small set Γ of wfs which could be checked for validity by model-theoretic methods and which would be such that every other wf valid in $\mathcal{V}$ could be obtained from it by some appropriate syntactic transformation.

One possible choice of Γ would involve as axioms all instances of the following three schemas:

1. $(A \rightarrow (B \rightarrow A))$
2. $((A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)))$
3. $(((\neg A) \rightarrow (\neg B)) \rightarrow (B \rightarrow A))$

where A, B, C range over the wfs in $\mathcal{F}$. Let us call this set of wfs $\Gamma_{\mathcal{V}}$. The wfs in $\Gamma_{\mathcal{V}}$ are valid in the frame $\mathcal{V}$ of all possible worlds and can be shown to axiomatize $\mathcal{V}$ in the sense that every wf valid in $\mathcal{V}$ can be 'proved' from these axioms in the manner described below. If we add a wf which is not provable from $\Gamma_{\mathcal{V}}$ to the set $\Gamma_{\mathcal{V}}$, the resultant set will be valid in a smaller set of worlds. Thus the enlarged set of axioms will be valid in a smaller frame. Let us look at an example. In the language with only two atoms, p and q, we add the wf p to the set $\Gamma_{\mathcal{V}}$. Which frame is determined by this new set of axioms? Certainly not the whole $\mathcal{V}$, because $f_3(p) = f_4(p) =$F. The frame that is determined by the new set of axioms is $\{f_1, f_2\}$ which is a proper subset of the set of all possible worlds, $\mathcal{V}$.

Now a wf B is *provable* from a set Γ (written as $\Gamma \vdash B$) iff there exists a sequence $A_1, A_2, \ldots, A_n$ of n wfs with the following properties:

- $A_n = B$ and
- each A_i is either an element of Γ (i.e. an axiom) or there exist prior members A_j, A_k of the sequence such that $A_k = A_j \rightarrow A_i$.

Intuitively, B is provable from Γ (i.e. $\Gamma \vdash B$) if B can be produced from members of Γ by applying the transformation rule

'given wfs C and $C \rightarrow D$, produce D'

zero, one or more times. This rule is known as 'modus ponens' or the 'rule of detachment' or '$\rightarrow$-elimination'. The sequence $A_1, A_2, \ldots, A_n$ is called a 'proof' or 'deduction' from Γ.

Let us look at two examples where a wf is proved from a set of wfs. First we take the set $\Gamma_{\mathcal{V}}$ and we try to deduce the wf $(p \rightarrow p)$. The following sequence of five wfs (written vertically so that we have room to justify the inclusion of each wf) is a proof from $\Gamma_{\mathcal{V}}$:

1.	$((p \to ((p \to p) \to p)) \to ((p \to (p \to p)) \to (p \to p)))$	(axiom with $A = p, B = (p \to p), C = p$)
2.	$(p \to ((p \to p) \to p))$	(axiom with $A = p, B = (p \to p)$)
3.	$((p \to (p \to p)) \to (p \to p))$	(modus ponens, from (1) and (2))
4.	$(p \to (p \to p))$	(axiom with $A = B = p$)
5.	$(p \to p)$	(modus ponens, from (3) and (4))

For the second example of a proof we add the set $\{(p \to q), (q \to r), p\}$ to the set $\Gamma_{\mathcal{V}}$. We assume we are working with the language where $\mathcal{P} = \{p, q, r\}$. Then

1.	$(p \to q)$	(axiom)
2.	$(q \to r)$	(axiom)
3.	p	(axiom)
4.	q	(modus ponens, from (1) and (3))
5.	r	(modus ponens, from (2) and (4)).

By the above we have shown that $\Gamma_{\mathcal{V}} \cup \{(p \to q), (q \to r), p\} \vdash r$.

The *deductive closure* $\mathrm{Th}(\Gamma)$ of a set Γ is the set of all wfs that are provable from Γ.

The proof-theoretic approach outlined above is *sound* in the following sense: for any set Γ of wfs where $\Gamma_{\mathcal{V}} \subseteq \Gamma$, if $\Gamma \vdash A$, then $\Gamma \models A$. So if A is provable from Γ then A is valid in the frame consisting of the models of Γ, thus $\mathrm{Th}(\Gamma) \subseteq \mathrm{Cn}(\Gamma)$.

The proof-theoretic approach symbolised by $\vdash$ is also *complete* in the sense that, for any set Γ of wfs where $\Gamma_{\mathcal{V}} \subseteq \Gamma$, if $\Gamma \models A$, then $\Gamma \vdash A$. So if A is valid in the frame consisting of the models of Γ, then A is provable from Γ, thus $\mathrm{Cn}(\Gamma) \subseteq \mathrm{Th}(\Gamma)$.

Say we want to determine whether $A \models B$ where A is arbitrary. Then the proof-theoretic approach outlined above can be adopted by taking as our set of axioms $\Gamma = \Gamma_{\mathcal{V}} \cup \{A\}$, in other words the wf A becomes an additional axiom and our frame becomes smaller. If we succeed in showing that $\Gamma \vdash B$, we may conclude that $A \models B$.

In conclusion, consider the limiting case of the empty frame $\mathcal{W} = \emptyset$. Can we axiomatise $\mathcal{W}$ in the sense of finding a set Γ of wfs such that every wf valid in $\mathcal{W}$ is provable from Γ? Well, every wf is valid in this (empty) frame, so one choice would be to set Γ equal to the set of all wfs. As noted earlier, though, this is unsatisfactory, because one prefers the set of axioms to be a relatively small subset of the set of valid wfs. Whatever set Γ we choose must have $\mathcal{W} = \emptyset$ as its set of models, so if a wf $A \in \Gamma$ is satisfiable in some world w, there must be another member of Γ, say B, which is not satisfiable in w. For example, B might be $\neg A$. The simplest such set, in the case of the language with $\mathcal{P} = \{p, q\}$, would be $\Gamma = \Gamma_{\mathcal{V}} \cup \{p, \neg p\}$. Given such contradictory axioms it is possible to deduce any wf A in the language. (To see this, note that $\neg p \to (p \to A)$ is valid in $\mathcal{V}$ for every wf A and is thus by completeness provable from $\Gamma_{\mathcal{V}}$. Given a proof of $\neg p \to (p \to A)$, modus ponens is then applied twice to yield A.) A set of axioms from which it is possible to deduce every wf is called *inconsistent*. It is a characteristic of the proof-theoretic approach outlined above that a set Γ is inconsistent iff it is unsatisfiable (i.e. has no models).

1.2 Predicate Logics

Predicate languages (first order languages) are more complicated but also more expressive than propositional languages, and therefore more useful for real-world applications. For simplicity's sake we do not include any function constants in the predicate languages that we are going to consider. We again need an alphabet $\mathcal{S}$ which must include the following symbols:

- for each positive n, zero, one or more predicate constants of 'arity' n, indicated by P_i^n,
- a countably infinite set of individual variables, indicated by x_i,
- zero, one or more individual constants, indicated by c_i,
- the connectives $\neg, \vee, \wedge, \rightarrow$ and $\leftrightarrow$,
- punctuation symbols (and) and
- quantifier symbols $\forall$ and $\exists$.

An example is the alphabet with one predicate constant, P_1^1, and two individual constants, c_1 and c_2.

Let $\mathcal{S}^*$ be the set of all possible strings over $\mathcal{S}$. The *terms* of the language are all the constants, c_i, and all the variables, x_i. The *atomic formulae* or *atoms* are defined as strings of the form $P_i^n(t_1, t_2, \ldots, t_n)$ where $t_1, t_2, \ldots, t_n$ are terms. Let $\mathcal{A}$ be the set of atoms. For example, the language with the alphabet given above would have $\mathcal{A} = \{P_1^1(c_1), P_1^1(c_2), P_1^1(x_1), P_1^1(x_2), \ldots\}$.

The *well-formed formulae* (wfs) of the predicate language are the members of the set $\mathcal{F}$ defined as follows:

$\mathcal{F}$ is the smallest subset of $\mathcal{S}^*$ such that

- $\mathcal{A} \subseteq \mathcal{F}$,
- if $A \in \mathcal{F}$, then $\neg A \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \vee B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \wedge B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \rightarrow B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \leftrightarrow B) \in \mathcal{F}$,
- if $A \in \mathcal{F}$, then $\forall x_i(A) \in \mathcal{F}$ and
- if $A \in \mathcal{F}$, then $\exists x_i(A) \in \mathcal{F}$.

In a wf $\forall x_i(A)$ we say that A is the *scope* of the quantifier and similarly for $\exists x_i(A)$. An occurrence of the variable x_i in a wf is said to be *bound* if it occurs within the scope of a $\forall x_i$ or a $\exists x_i$ or if it is the x_i next to the quantifier symbol. If an occurrence of a variable is not bound, it is said to be *free*. A wf with no free variables is a *sentence*. (We see that

here, in contrast to propositional languages, there are wfs that are not sentences.) An atom which is a sentence is a *ground* atom.

Let us look at a few examples. In the wf $\exists x_2(P_1^2(x_1, x_2))$ the scope of the quantifier $\exists x_2$ is $P_1^2(x_1, x_2)$ so that both occurrences of x_2 in $\exists x_2(P_1^2(x_1, x_2))$ are bound, but the occurrence of x_1 is free. An example of a sentence is $\forall x_1(\exists x_2(P_1^2(x_1, x_2)))$. An example of a ground atom is $P_1^2(c_3, c_5)$.

In the rest of this section, we follow the substitution approach to quantification as explained by Shoenfield [Shoenfield 1967]. We start by defining substitution instances of wfs.

Let A be any member of $\mathcal{F}$. For every variable x and individual constant c the formula A_c^x is given by the following set of rules.

1. If $A \in \mathcal{A}$, then A_c^x is the result of substituting c for every occurrence of x in A.

2. If $A \in \mathcal{F}$, then $(\neg A)_c^x = \neg A_c^x$.

3. If $A, B \in \mathcal{F}$, then $(A \odot B)_c^x = (A_c^x \odot B_c^x)$ where $\odot \in \{\vee, \wedge, \rightarrow, \leftrightarrow\}$.

4. If $A \in \mathcal{F}$, then $(Qx(A))_c^x = Qx(A)$ where $Q \in \{\forall, \exists\}$.

5. If $A \in \mathcal{F}$, then $(Qy(A))_c^x = Qy(A_c^x)$ where $Q \in \{\forall, \exists\}$ and $x \neq y$.

More generally, if ψ maps terms t_1 to s_1, t_2 to s_2, ..., t_n to s_n, then $A^\psi = (((A_{s_1}^{t_1})_{s_2}^{t_2}) \ldots {}_{s_n}^{t_n})$.

Let U be any nonempty set, the *universe of discourse*. An example is the set $\{Ctn, Jhb, Dbn, Pta\}$. Now add the elements of U to the set of individual constants to form the set U'. Say we specify our language as having two individual constants, c_1 and c_2, and one predicate constant, P_1^2. Then we will have $U' = \{Ctn, Jhb, Dbn, Pta, c_1, c_2\}$.)

A *U-formula* is then defined exactly like a well-formed formula except that the elements of U' may also be used as terms. A *U-atom*, however, is defined as a ground atom containing only elements of U. In the example above the set of U-atoms will be

$\{P_1^2(Ctn, Ctn), P_1^2(Ctn, Jhb),$
$P_1^2(Ctn, Dbn), P_1^2(Ctn, Pta),$
$P_1^2(Jhb, Ctn), P_1^2(Jhb, Jhb),$
$P_1^2(Jhb, Dbn), P_1^2(Jhb, Pta),$
$P_1^2(Dbn, Ctn), P_1^2(Dbn, Jhb),$
$P_1^2(Dbn, Dbn), P_1^2(Dbn, Pta),$
$P_1^2(Pta, Ctn), P_1^2(Pta, Jhb),$
$P_1^2(Pta, Dbn), P_1^2(Pta, Pta)\}$.

Now an *assignment* is defined as an assignment of truth values to all U-atoms, i.e. f is a function from the set of U-atoms to the set {T,F}. In the above example we could specify that $f(P_1^2(Jhb, Ctn)) = f(P_1^2(Jhb, Dbn)) = f(P_1^2(Jhb, Pta))$ =T and further specify that truth value F is assigned to all other U-atoms.

An *interpretation* (or world, or situation, or state) is a triple $w = (\phi, f, U)$ where U is the universe of discourse, ϕ is a mapping from the individual constants of the language to the universe of discourse and f is an assignment. In the above example we could specify $\phi(c_1) = Jhb$ and $\phi(c_2) = Ctn$. The purpose of the mapping ϕ is to allow us to go from an 'ordinary' atom to a U-atom.

Let $w \models A$ abbreviate the assertion that the sentence A has truth value T under the interpretation w (or in the world or situation or state w). Similarly, by $w \not\models A$ is meant that the sentence A has truth value F under the interpretation w (or in the world w). Then truth values are given to the sentences of our language by the following rules:

- $w \models A$ iff A is a ground atom and $f(A^\phi)$=T where A^ϕ is the string in which every constant c in A has been replaced by $\phi(c)$, i.e. A^ϕ is the U-atom corresponding, under the mapping ϕ, to the original atom A,
- $w \models \neg A$ iff $w \not\models A$,
- $w \models (A \vee B)$ iff $w \models A$ or $w \models B$,
- $w \models (A \wedge B)$ iff $w \models A$ and $w \models B$,
- $w \models (A \rightarrow B)$ iff $w \not\models A$ or $w \models B$,
- $w \models (A \leftrightarrow B)$ iff $w \models A$ and $w \models B$, or $w \not\models A$ and $w \not\models B$,
- $w \models \forall x(A)$ iff for all $k \in U$, $w \models (A_k^x)$ and
- $w \models \exists x(A)$ iff for some $k \in U$, $w \models (A_k^x)$.

If a sentence A has truth value T, i.e. if $w \models A$, we say that A is 'true' in the world w and that w is a 'model' of the sentence A. Similarly, if $w \not\models A$ we say that the sentence A is 'false' in the world w and that w is not a model of the sentence A. In the world as specified in the above example the sentence $P_1^2(c_2, c_1)$ will be false because $f(P_1^2(Ctn, Jhb))$ =F, so w is not a model of this sentence. It is, however, a model of the sentence $\exists x_1(P_1^2(c_1, x_1))$.

Let $\mathcal{W}$ be a set of interpretations (worlds) and let us call such a set a *frame*. Then we say that a sentence A is *satisfiable in* $\mathcal{W}$ if it is true under at least one interpretation in $\mathcal{W}$ and that it is *valid in the frame* $\mathcal{W}$ if it is true under all interpretations in the frame $\mathcal{W}$.

A set Γ of sentences is satisfiable in $\mathcal{W}$ if every sentence in Γ is true in at least one world of $\mathcal{W}$, i.e. if at least one interpretation in $\mathcal{W}$ is a model of every sentence in the set. The set Γ of sentences is valid in the frame $\mathcal{W}$ if each sentence in the set is true in every world in $\mathcal{W}$, i.e. if each world in the frame is a model of every sentence in Γ.

Let us look at an example using the language which has one predicate constant P_1^1 and two individual constants c_1 and c_2. We take as our universe of discourse $U = \{10, 13, 15\}$. Then the set of U-atoms is $\{P_1^1(10), P_1^1(13), P_1^1(15)\}$. We further specify the mapping ϕ as follows: $\phi(c_1) = 13$ and $\phi(c_2) = 10$. Let the assignment f in the interpretation $w_1 = \{\phi, f, U\}$ be defined to be

$f(P_1^1(10))$ =T, $f(P_1^1(13))$ =F, $f(P_1^1(15))$ =F.

(We think of $P_1^1(x)$ as 'x is even'.) Is the ground atom $A = P_1^1(c_1)$ true in this world? We know that $f(A^\phi) = f(P_1^1(13))$ =F, so this ground atom is not true in world w_1.

Is the sentence $B = \exists x_1(P_1^1(x_1))$ true in the world w_1? Well, we know that $f(P_1^1(10)) =$ T, so we have that $w_1 \models P_1^1(x_1)_{10}^{x_1}$, so we do have that $w_1 \models \exists x_1(P_1^1(x_1))$. The sentence $\forall x_1 P_1^1(x_1)$, however, is not true in the world w_1, because $f(P_1^1(13)) = f(P_1^1(15))$ =F.

Say we define a world w_2 identical to w_1 except for the assignment f which is now specified to be

$f(P_1^1(10))$ =F, $f(P_1^1(13))$ =T, $f(P_1^1(15))$ =F.

(We think of $P_1^1(x)$ as 'x is prime'.) Again the sentence $B = \exists x_1(P_1^1(x_1))$ is true (in the world w_2) and the sentence $\forall x_1 P_1^1(x_1)$ is not. This means that the sentence $B =$

$\exists x_1(P_1^1(x_1))$ is valid in the frame $\{w_1, w_2\}$, i.e. every world in this frame is a model of the sentence B.

As with propositional logic, every sentence, and indeed every set of sentences, determines a particular frame, namely the set of worlds that are models of the specific sentence or set of sentences. If a sentence A is valid in the frame determined by the set of sentences Γ, we write $\Gamma \models A$. A simple extension of the proof-theoretic approach sketched in section 1.1.3 provides a notion of proof that is sound and complete, i.e. is such that, for all sets Γ of wfs, $\Gamma \models A$ iff $\Gamma \vdash A$, or equivalently, $\mathrm{Cn}(\Gamma) = \mathrm{Th}(\Gamma)$.

1.3 Propositional Modal Logics

In the above sections we have considered the syntax and semantics of classical propositional and predicate logics. Now we introduce *modal operators* and then look at the possible world semantics.

1.3.1 Modal Operators

When we define a language we need an alphabet. We use the same set of symbols that served for propositional logic, but we add (at least) one member to the set namely the box $\Box$. In the literature a wide variety of symbols are used for this additional member(s), for example K_a [Fitting 1993], K_a and B_a [Hintikka 1962], L [Lukaszewicz 1990], $[S_i]$ [Konolige 1986].

Let $\mathcal{P}$ be the set of atoms. Then the *alphabet* of the modal language is defined to be the set

$\mathcal{S} = \mathcal{P} \cup \{\neg, \vee, \wedge, \rightarrow, \leftrightarrow, \Box, (,)\}$.

Let $\mathcal{S}^*$ be the set of all possible strings over $\mathcal{S}$. The *well-formed formulae* (wfs) or *sentences* of the propositional modal language are the members of the set $\mathcal{F}$ defined as follows:

$\mathcal{F}$ is the smallest subset of $\mathcal{S}^*$ such that

- $\mathcal{P} \subseteq \mathcal{F}$,
- if $A \in \mathcal{F}$, then $\neg A \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \vee B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \wedge B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \rightarrow B) \in \mathcal{F}$,
- if $A, B \in \mathcal{F}$, then $(A \leftrightarrow B) \in \mathcal{F}$ and
- if $A \in \mathcal{F}$, then $\Box(A) \in \mathcal{F}$.

As always, we omit parentheses if no ambiguity can arise.

Examples of wfs in a modal language having $\mathcal{P} = \{p, q\}$ as set of atoms are
$(\neg p \rightarrow \Box(p))$
$((p \vee q) \rightarrow (q \vee p))$
$(\Box(p) \leftrightarrow \Box(\Box(p)))$ or, omitting some parentheses, $\Box p \leftrightarrow \Box(\Box p)$.

1.3.2 Possible World Semantics of Propositional Modal Languages

Modal logics are logics of qualified truth [Fitting 1993]. Let $A \in \mathcal{F}$. There are a number of ways in which $\Box A$ can be read. Here are a few [Goldblatt 1992]:

- It is necessarily true that A.
- It will always be true that A.
- It ought to be that A.
- It is known that A.
- It is believed that A.
- After the program terminates, A.

We have a reasonably clear idea of the circumstances under which a nonmodal wf A is true in a world w. But under what circumstances may we claim that a wf of the form $\Box A$ is true in a world w? To a degree the answer depends on the way $\Box$ is intended to be read, but we make some general remarks.

When we discussed the truth of nonmodal propositional sentences and nonmodal predicate languages we found it useful to use the notions 'world' and 'frame'. We also gave precise definitions of these notions. In a propositional language worlds may be identified with assignments of truth values to atoms. In a predicate language, however, many different worlds (or interpretations) may give rise to the same assignment of truth values to atomic sentences. Let us look at an example. Consider the predicate language with two individual constants c_1 and c_2 and one predicate constant P_1^2. Consider the worlds w_1 and w_2 defined as follows:

$w_1 = \{\phi, f, \{0,1\}\}$ with $\phi(c_1) = 0, \phi(c_2) = 1$ and
$f(P_1^2(0,1))$ =T, $f(P_1^2(0,0)) = f(P_1^2(1,1)) = f(P_1^2(1,0))$ =F,
$w_2 = \{\phi, f, \{1,2\}\}$ with $\phi(c_1) = 1, \phi(c_2) = 2$ and
$f(P_1^2(1,2))$ =T, $f(P_1^2(1,1)) = f(P_1^2(2,2)) = f(P_1^2(2,1))$ =F.

In each of the worlds w_1 and w_2, the atomic sentence $P_1^2(c_1, c_2)$ is true while the atomic sentences $P_1^2(c_1, c_1)$, $P_1^2(c_2, c_2)$ and $P_1^2(c_2, c_1)$ are false.

In the semantics of modal languages the notions of world and frame are adapted as follows. The truth of a wf will be assessed in terms of a structure outside the language which will be called a 'model'. This terminology differs from the previous use of the word and is, of course, unfortunate but traditional. This 'model' will consist of a 'frame' and a 'valuation'. A frame will be, as before, a set of possible worlds but will have an additional feature: the worlds may be related to each other in a nontrivial way. The worlds themselves need not necessarily be assignments or interpretations although, of course, there is nothing that prevents us from choosing the worlds in such a way. We should think of them as abstract points or labels to which we may attribute some concrete meaning. An example is the case of temporal logic where the worlds are taken to be instants in time. A frame, then, is a pair $\langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is some set of worlds and $\mathcal{R}$ is some relation on $\mathcal{W}$, for example $\mathcal{R}$ may be the identity relation on $\mathcal{W}$, or the empty relation on $\mathcal{W}$, or some sort of order relation on $\mathcal{W}$ (perhaps expressing temporal succession).

The second part of a model, namely a valuation, serves to connect the members of $\mathcal{W}$ with assignments of truth values to atoms: for each world w in $\mathcal{W}$, the valuation will specify which atoms are true in w.

To summarise: A *frame* is defined as a pair $\langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{W}$ is a non-empty set and $\mathcal{R}$ is a binary relation on $\mathcal{W}$. The relation $\mathcal{R}$ is called an *accessibility relation*. For $w_1, w_2 \in \mathcal{W}$, we say 'world w_2 is *accessible* from world w_1' iff $(w_1, w_2) \in \mathcal{R}$.

A *valuation* in a frame $\langle \mathcal{W}, \mathcal{R} \rangle$ is a function v with domain $\mathcal{W} \times \mathcal{P}$ and range {T,F}. Recall that $\mathcal{P}$ is the set of atoms. Intuitively $v(w,p)$=T means that p is true at the world $w \in \mathcal{W}$.

A *model* is a triple $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ where $\langle \mathcal{W}, \mathcal{R} \rangle$ is a frame and v is a valuation in it. We call a model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ *explicit* iff $\mathcal{W} \subseteq \mathcal{V}$, the set of assignments of truth values to the atomic sentences of the language, and v is defined in the obvious way namely $v(w,p) = w(p)$ for all $w \in \mathcal{W}$ and $p \in \mathcal{P}$.

Once truth values are assigned to atoms, truth values of all formulae can be found. We write $\mathcal{M} \mid\!\vdash_w A$ with the intended meaning of 'The wf A is true at world w of the model $\mathcal{M}$.' We write $\mathcal{M} \mid\!\nvdash_w A$ with the intended meaning of 'The wf A is not true, i.e. is false, at the world w in the model $\mathcal{M}$.' So, let $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ be a model. Then the following holds:

$\mathcal{M} \mid\!\vdash_w p$	iff	$v(w,p) = \mathrm{T}$ where $p \in \mathcal{P}$
$\mathcal{M} \mid\!\vdash_w \neg B$	iff	$\mathcal{M} \mid\!\nvdash_w B$
$\mathcal{M} \mid\!\vdash_w (B \vee C)$	iff	$\mathcal{M} \mid\!\vdash_w B$ or $\mathcal{M} \mid\!\vdash_w C$
$\mathcal{M} \mid\!\vdash_w (B \wedge C)$	iff	$\mathcal{M} \mid\!\vdash_w B$ and $\mathcal{M} \mid\!\vdash_w C$
$\mathcal{M} \mid\!\vdash_w (B \rightarrow C)$	iff	$\mathcal{M} \mid\!\vdash_w C$ or $\mathcal{M} \mid\!\nvdash_w B$
$\mathcal{M} \mid\!\vdash_w (B \leftrightarrow C)$	iff	either $\mathcal{M} \mid\!\vdash_w B$ and also $\mathcal{M} \mid\!\vdash_w C$,
		or $\mathcal{M} \mid\!\nvdash_w B$ and also $\mathcal{M} \mid\!\nvdash_w C$
$\mathcal{M} \mid\!\vdash_w \Box B$	iff	$\mathcal{M} \mid\!\vdash_{w'} B$ for all worlds $w' \in \mathcal{W}$ which are accessible from w.

The last item above specifies that a wf $\Box A$ is true at a world w in a model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ iff the wf A is true at all worlds w' such that $(w, w') \in \mathcal{R}$. This is one useful way to give meaning to $\Box$. Other ways will be described in later chapters.

Let us look at an example. Say we have the language with $\mathcal{P} = \{p, q\}$ and we consider the model
$\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle = \langle \{w_1, w_2, w_3, w_4\}, \{(w_1, w_2), (w_1, w_3), (w_1, w_4), (w_2, w_2), (w_2, w_4)\}, v \rangle$
where the function v is defined as follows:
$v(w_1, p)$ =T, $v(w_2, p)$ =T, $v(w_3, p)$ =F, $v(w_4, p)$ =F,
$v(w_1, q)$ =T, $v(w_2, q)$ =F, $v(w_3, q)$ =T, $v(w_4, q)$ =F.
Is the wf $A = \Box(p \vee \neg q)$ true at world w_1? Well, the wf $(p \vee \neg q)$ is true at worlds w_1, w_2 and w_4, but it is false at world w_3. Because worlds w_2, w_3 and w_4 are all accessible from world w_1 and $(p \vee \neg q)$ is not true at all these worlds, the wf $\Box(p \vee \neg q)$ is not true at world w_1. Note, however, that $\Box(p \vee \neg q)$ is true at world w_2 because $(p \vee \neg q)$ is true at the two worlds accessible from world w_2, namely w_2 itself and w_4. Further note that $\Box(p \vee \neg q)$ is also vacuously true at worlds w_3 and w_4 – no world is accessible from either of them.

Let us keep the model as above and consider the wf $\Box(\Box(p \vee \neg q))$. Is this wf true at world w_1? Well, we have already seen that the wf $\Box(p \vee \neg q)$ is true at worlds w_2, w_3 and w_4, i.e. is true at all three worlds accessible from w_1, so the wf $\Box(\Box(p \vee \neg q))$ is true at world w_1. The wf $\Box(\Box(p \vee \neg q))$ is also true at world w_2 because (as we have seen above) the wf $\Box(p \vee \neg q)$ is true at the two worlds accessible from w_2, namely at worlds w_2 and w_4. Furthermore, the wf $\Box(\Box(p \vee \neg q))$ is vacuously true at worlds w_3 and w_4 because no world is accessible from these two worlds. This means that the wf $\Box(\Box(p \vee \neg q))$ is true at all four worlds of the frame of this model.

A wf A which is true at all words w of the model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$, such as the last example above, is said to be *valid (*or *globally true) in the model* $\mathcal{M}$ and we write it as $\mathcal{M} \Vdash A$. Assume that some set of models is chosen (for example, the set of all models $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ associated with the frame $\langle \mathcal{W}, \mathcal{R} \rangle$). Relative to this semantics, it is possible to define notions of entailment and provability, and to examine the relationship between the notions.

A wf A is entailed by a set Γ of wfs iff A is valid in every model (of the chosen semantics) in which each member of Γ is valid. We write it as $\Gamma \models_{pw} A$ where the pw serves to remind us that truth is defined in terms of possible worlds. $\mathrm{Cn}(\Gamma)$ is the set of all wfs entailed by Γ. As before we can construct a proof architecture and write $\Gamma \vdash A$ if A is provable from the set Γ. The set of all wfs provable from Γ is indicated by $\mathrm{Th}(\Gamma)$. The approach will be sound iff $\mathrm{Th}(\Gamma) \subseteq \mathrm{Cn}(\Gamma)$ and it will be complete iff $\mathrm{Cn}(\Gamma) \subseteq \mathrm{Th}(\Gamma)$.

1.4 Logics of Belief

Given a modal propositional language, how well can the knowledge or beliefs of an agent (or several agents) be represented in it? In order to read $\Box A$ as 'The agent knows A' or 'The agent believes A', what restrictions or adaptations to the semantics would be necessary? What properties do we expect the accessibility relation to have? Let us investigate this last question.

Consider any complex system that needs to be controlled. Examples are a nuclear powerplant, a chemical factory and a submarine. We think of such a system as having a control room manned by an agent. The control room is separate from the system but receives information about the system via sensors that monitor (some of the) components. A state of the system will be represented by an assignment f for a language with an appropriate set of atoms, say $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$. The information available to the agent in the control room would give the values $f(p_i)$ for (usually) some of the atoms. Without loss of generality we may assume that the control room reveals the values of $f(p_1), f(p_2), \ldots, f(p_k)$ for some $k \leq n$.

Let $\mathcal{V}$ be the set of all possible assignments $f : \mathcal{P} \to \{\mathrm{T},\mathrm{F}\}$. Now assume that the system is in the state represented by assignment f. The only information available to the agent is that which is provided by the control room. As far as the agent is able to say, therefore, the system may be in any state that is represented by an assignment f' such that $f'(p_1) = f(p_1)$, $f'(p_2) = f(p_2)$, ..., $f'(p_k) = f(p_k)$. In other words the set $\mathcal{V}$ is partitioned into disjoint nonempty subsets, each of which contains the assignments which agree on $p_1, p_2, \ldots, p_k$. (If, for example, $n = 7$ and $k = 3$ we will have eight equivalence classes, each containing sixteen assignments.) If $k = 0$ the partition consists of a single subset namely $\mathcal{V}$ since all assignments agree on $p_1, p_2, \ldots, p_k$ for this value of k. If $k = n$ the partition consists of the singleton subsets of $\mathcal{V}$ since no two assignments can agree on all p_i. We see that the agent's knowledge (or beliefs) may be represented with the help of an equivalence relation on $\mathcal{V}$. The more limited the information provided by the control room, i.e. the closer k is to 0, the closer the equivalence relation is to (the trivial) $\mathcal{V} \times \mathcal{V}$. The more complete the available information, i.e. the closer k is to n, the closer the equivalence relation is to the identity relation on $\mathcal{V}$.

Let $\mathcal{M} = \langle \mathcal{V}, \mathcal{R}, v \rangle$ where $\mathcal{R}$ is the equivalence relation induced by such a partition on $\mathcal{V}$ and v the obvious valuation given by $v(f, p_i) = f(p_i)$ for each $i \leq n$. In terms of our earlier definition $\mathcal{M} \Vdash_f \Box A$ iff $\mathcal{M} \Vdash_{f'} A$ for every assignment f' such that $(f, f') \in \mathcal{R}$. This

expresses the intuition that when the system is in the state (represented by) f, the agent, being unable to distinguish between f and the members f' of the set $\{f' \mid (f, f') \in \mathcal{R}\}$, will believe those wfs A that hold in all the equivalent states f'. For example let $k < n$ and let the system be in state f with $f(p_i)$ =T for all $i \leq n$. Then the agent will believe the wf $p_1 \wedge p_2 \wedge \ldots \wedge p_k$, i.e. it will be the case that $\mathcal{M} \mid\!\vdash_f \Box(p_1 \wedge p_2 \wedge \ldots \wedge p_k)$. The agent, however, will not believe $p_1 \wedge p_2 \wedge \ldots \wedge p_k \wedge p_{k+1}$ even though this wf is true in the state f because there exist assignments f' which are equivalent to f but where this wf is false. Formally, $\mathcal{M} \not\mid\!\vdash_f \Box(p_1 \wedge p_2 \wedge \ldots \wedge p_k \wedge p_{k+1})$ because there exists a state represented by f' such that $(f, f') \in \mathcal{R}$ and $\mathcal{M} \not\mid\!\vdash_{f'} (p_1 \wedge p_2 \wedge \ldots \wedge p_k \wedge p_{k+1})$. We could just take the assignment f' such that $f'(p_i) = f(p_i)$ for all $i \neq k+1$ and $f'(p_{k+1})$=F.

Thus, given a model and a specific world, an agent knows or believes A if A is true in all worlds which he considers may be the actual state of the system, based on the information that he has. (Note that the use of the masculine singular pronoun is generic and not gendered.)

1.4.1 One Ideally Reasoning Agent

The above example shows that there may be situations where it is appropriate to choose the accessibility relation $\mathcal{R}$ to be an equivalence relation on the set $\mathcal{W}$ of worlds. As we know, an equivalence relation is reflexive, symmetric and transitive. What does this mean in terms of an agent and his beliefs?

Suppose $\mathcal{R}$ is reflexive, i.e. the pair (w, w) is a member of $\mathcal{R}$ for every $w \in \mathcal{W}$. This means that if the actual situation is w, the information available to the agent is such that he does not exclude the possibility that world w is the 'real' state of the system. In the example of the agent in the control room of a nuclear powerplant the real state w will be represented by an assignment agreeing on the atoms $p_1, p_2, \ldots, p_k$ with the information that is available to the agent. Thus the sensors are functioning correctly and the agent believes the information provided to him.

For symmetry we have that for all $w, w' \in \mathcal{W}$, if $(w, w') \in \mathcal{R}$, then also $(w', w) \in \mathcal{R}$. Suppose the actual state of affairs is world w and the agent cannot distinguish between worlds w and w'. (His information is such that he considers w' as possibly the case while the situation actually is w.) Symmetry implies that if the actual situation were w', the information available to him would be such that he would consider world w as possibly the case. In the above example it means that the assignments representing states w and w' agree on atoms $p_1, p_2, \ldots, p_k$ and if either of the two states is the actual state of the system, the agent will include the other in the set of states that he considers possible.

Assume the relation is transitive, i.e. for all $w, w', w'' \in \mathcal{W}$, if (w, w'), $(w', w'') \in \mathcal{R}$, then $(w, w'') \in \mathcal{R}$. What does this mean for our agent? Well, if the agent cannot distinguish between worlds w and w' when the actual state of affairs is w, and also cannot distinguish between worlds w' and w'' when the actual state of affairs is w', transitivity suggests he will not be able to distinguish between states w and w'' if the actual situation is w. Consider again the agent in the control room of the nuclear powerplant. If the assignments representing states w and w' agree on atoms $p_1, p_2, \ldots, p_k$ and those representing states w' and w'' agree on $p_1, p_2, \ldots, p_k$, then the assignments representing states w and w'' also agree on $p_1, p_2, \ldots, p_k$.

Suppose we restrict the accessibility relation to be an equivalence relation. What kind of beliefs will an agent have in such a case? It may be of help if we answer the following question: Which wfs will be valid in such frames?

First of all, consider a nonmodal wf A. We saw examples of such wfs that were valid in the nonmodal frame $\mathcal{V}$ consisting of all the truth assignments $f : \mathcal{P} \rightarrow \{\text{T,F}\}$. An example of such a wf is $p \vee \neg p$. We can easily show that such wfs are also valid in every modal frame $\langle \mathcal{W}, \mathcal{R} \rangle$. We do it as follows:

Suppose $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ is any model relative to the frame $\langle \mathcal{W}, \mathcal{R} \rangle$ and A is any nonmodal wf valid in the nonmodal frame $\mathcal{V}$. Then we have, for all $w \in \mathcal{W}$, that $\mathcal{M} \mid\!\vdash_w A$. Why? The reason is that the truth of (the nonmodal) A at w is determined in the usual manner by the valuation that extends the truth assignment $f \in \mathcal{V}$ defined by $f(p) = v(w, p)$ for every $p \in \mathcal{P}$. Note that the argument is independent of whether $\mathcal{R}$ is an equivalence relation or not. So the result holds for all accessibility relations.

What can be said of the validity of wfs involving $\Box$ in a frame with $\mathcal{R}$ an equivalence relation? It has been shown in the literature, for example [Goldblatt 1992], that given certain properties of a relation, certain schemas are valid in the corresponding frames, and vice versa. For example, when the accessibility relation of a frame is reflexive the schema $\Box A \rightarrow A$ is valid in that frame (see below), and conversely, if the schema $\Box A \rightarrow A$ is valid in a frame the accessibility relation of the frame is reflexive.

In order to get a clearer picture of an agent's beliefs we select the following four modal schemas which are known to characterise frames $\langle \mathcal{W}, \mathcal{R} \rangle$ where $\mathcal{R}$ is an equivalence relation:

1. $(\Box A \wedge \Box(A \rightarrow B)) \rightarrow \Box B$
2. $\Box A \rightarrow A$
3. $\Box A \rightarrow \Box\Box A$
4. $\neg\Box A \rightarrow \Box\neg\Box A$

As an exercise in applying the concepts of section 1.3 we show that all instances of these schemas are valid in such frames. Let $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ be a model with $\mathcal{R}$ an equivalence relation and let $w \in \mathcal{W}$.

A wf of the form $(\Box A \wedge \Box(A \rightarrow B)) \rightarrow \Box B$ can only be false at world w if $\Box B$ is false at w but $\Box A \wedge \Box(A \rightarrow B)$ is true at w. Suppose this is the case. This means that the wf B is false at at least one world accessible from w, say at w'. For the wf $\Box A \wedge \Box(A \rightarrow B)$ to be true at w, both $\Box A$ and $\Box(A \rightarrow B)$ must be true at w. So the wf A must be true at all worlds accessible from w, thus also at w'. Similarly the wf $A \rightarrow B$ must be true at all worlds accessible from w, thus also at w'. Since A is true and B is false at w' this is, however, impossible. We have thus shown that this (first) schema is valid. Note that it is valid in any frame; it does not depend on the properties of the accessibility relation.

A wf of the form $\Box A \rightarrow A$ can only be false at world w if $\Box A$ is true at w but A is false at w. For $\Box A$ to be true at w, A must be true at all worlds accessible from w, so also at w itself because of reflexivity. So $\Box A \rightarrow A$ cannot be false at w.

Suppose a wf of the form $\Box A \rightarrow \Box\Box A$ is false at world w. This can only be the case if $\Box A$ is true but $\Box\Box A$ is false at w. For $\Box\Box A$ to be false at w, the wf $\Box A$ must be false at at least one world accessible from w, say at world w'. This means that the wf A must be false at at least one world accessible from w', say at world w''. On the other hand, for the wf $\Box A$ on the lefthand side of the original wf to be true at w, A must be true at both world w' and world w'' (which are both accessible from w, because of transitivity). Since A cannot be both true and false at w'' this is not possible.

Finally a wf of the form $\neg\Box A \rightarrow \Box\neg\Box A$ can only be false at world w if $\neg\Box A$ is true at w but $\Box\neg\Box A$ is false at w. Suppose this is the case. For the wf $\neg\Box A$ to be true at world

w, the wf $\Box A$ must be false at w, i.e. A must be false at some world accessible from w, say world w'. For the righthand side wf to be false at w, the wf $\neg\Box A$ must be false at a world accessible from w, say at w''. This means the wf $\Box A$ must be true at w'' and this means that the wf A must be true at all worlds accessible from w''. Which worlds are accessible from w''? Well, we know that $(w, w''), (w'', w), (w, w') \in \mathcal{R}$ – remember the relation is symmetric – and so also $(w'', w') \in \mathcal{R}$ because of transitivity. This then leads to a contradiction: A is both true and false at world w'.

We have now shown that the four given modal schemas are valid in all frames with $\mathcal{R}$ an equivalence relation. All wfs which are valid in such frames can be deduced from these four axiom schemas and the following two rules (see for example [Fagin et al 1995]):

from A and $A \rightarrow B$ infer B (modus ponens) and

from A infer $\Box A$ (necessitation).

What do the four modal axiom schemas suggest about the beliefs or knowledge of an agent, given a model and a specific world?

The fact that the agent knows or believes a wf A when the system is in state w is expressed by $\mathcal{M} \mid\!\vdash_w \Box A$. Since the second axiom schema, $\Box A \rightarrow A$, is valid, it follows that $\mathcal{M} \mid\!\vdash_w \Box A \rightarrow A$. Thus $\mathcal{M} \mid\!\vdash_w A$, i.e. A really is true in state w. The axiom schema $\Box A \rightarrow A$ represents the notion that the agent knows (believes) only wfs that are true. This is usually taken as the key axiom when making a distinction between knowledge and belief. It seems feasible to have such an axiom in logics of knowledge, i.e. the agent can only *know* something if that something is true. In logics of belief, however, this is too strict a constraint – it is quite possible for an agent to believe something that is actually not true.

Suppose the system is in state w and the agent knows or believes A, i.e. suppose $\mathcal{M} \mid\!\vdash_w \Box A$. Since the schema $\Box A \rightarrow \Box\Box A$ is valid, $\mathcal{M} \mid\!\vdash_w \Box A \rightarrow \Box\Box A$, thus $\mathcal{M} \mid\!\vdash_w \Box\Box A$ which means that the agent knows that he knows A. The third schema above therefore represents the notion that the agent is able to perform positive introspection. Similarly, suppose the agent does not know or believe A when the system is in state w, i.e. suppose $\mathcal{M} \mid\!\vdash_w \neg\Box A$. Since the schema $\neg\Box A \rightarrow \Box\neg\Box A$ is valid, $\mathcal{M} \mid\!\vdash_w \neg\Box A \rightarrow \Box\neg\Box A$, thus $\mathcal{M} \mid\!\vdash_w \Box\neg\Box A$, which means that the agent knows that he does not know A. So the fourth schema above represents the fact that the agent is able to perform negative introspection. Intuitively these two axiom schemas say that the agent is able to look at his knowledge base and be conscious of everything he knows and of everything he does not know.

The schema $(\Box A \wedge \Box(A \rightarrow B)) \rightarrow \Box B$ together with the rule of necessitation (from A infer $\Box A$) ensures that the agent is *logically omniscient*. For example, necessitation expresses the fact that if $\mathcal{M} \mid\!\vdash A$, then $\mathcal{M} \mid\!\vdash \Box A$. In other words, if A is valid then the agent knows A. Furthermore, suppose the agent knows A and also knows $A \rightarrow B$. One would then expect a suitably gifted agent to know B, and this is guaranteed by the first axiom schema, namely if $\mathcal{M} \mid\!\vdash_w \Box A$ and $\mathcal{M} \mid\!\vdash_w \Box(A \rightarrow B)$, then $\mathcal{M} \mid\!\vdash_w \Box B$.

In this section we have therefore described a situation where an agent has knowledge (rather than beliefs which may be mistaken), knows what he knows and what he does not know, and further knows all the consequences of the things that he does know. Is this a good approximation of the situation in 'real life'? We will investigate the question in the next section.

1.4.2 Points of Criticism

In the previous section we have assumed that the accessibility relation $\mathcal{R}$ of the frames we considered were equivalence relations. This, however, is not always realistic.

Let us consider the agent in the control room of a nuclear powerplant and suppose it is possible that some of the sensors are malfunctioning. Now we can no longer assume that the available information will induce a partition on the set $\mathcal{W}$ of possible worlds. Say, for example, the sensor providing the information necessary to get $f(p_1)$ does not work correctly. The agent (not knowing this) believes the evidence and does not consider the actual state of the powerplant to be a possible state. Thus the accessibility relation is no longer reflexive.

From the above it is clear that it is not realistic to assume that a partition is always induced by the information available to the agent, i.e. it is not always realistic to assume that the accessibilty relations of the frames are equivalence relations.

What properties will the knowledge or beliefs of an agent have if we do not require $\mathcal{R}$ to be an equivalence relation? Well, suppose we consider frames where we do not require the accessibility relation to be reflexive. Then the second axiom schema above will fail to hold, i.e. not all wfs of the form $\Box A \rightarrow A$ will be valid in these frames. This means then that the restriction that an agent can only know something if it is true falls away. In those cases it is more appropriate to speak of believe instead of know. If we do not require the accessibility relation to be reflexive, we can include the situation where an agent believes something that is actually not true.

If we do not require the accessibility relation to be symmetric, negative introspection falls away because then the fourth axiom schema above will fail to hold. This corresponds to a situation where the agent is not aware of everything that he does not believe. If we do not require transitivity the third and fourth axiom schemas fail to hold and both positive and negative introspection fall away. This corresponds to a situation where the agent is not always aware of what he believes.

It is clear that by changing the properties of the accessibility relation we can remove positive or negative introspection and also the restriction that only true things are believed. However, the logical omniscience expressed by the first axiom schema and necessitation is a consequence of the possible world semantics itself and is unaffected by the properties of the accessibility relation $\mathcal{R}$. In order to deal with more realistic agents it will be necessary to find an alternative to possible world semantics.

1.5 Summary

In this chapter we have briefly looked at the syntax and semantics of classical non-modal propositional languages and of (in less detail) classical non-modal predicate languages. In the final part of the chapter we have considered classical propositional modal languages – the syntax and some aspects of possible world semantics. We have, in particular, investigated the possible world semantics when the modal operator is read as 'The agent knows ...' or 'The agent believes ...'.

In the following chapters we are going to take a look at different ways in which authors handle modal logics when the meaning of belief is attached to the modal operator. Some authors consider propositional modal logics only, while others look at both propositional and predicate modal logics. We will handle the latter as and when required. In the final chapter of the dissertation we take a brief look at guidelines for changing a set of beliefs when new information becomes available.

Chapter 2

Moore's Autoepistemic Logic

There are four sorts of men:
he who knows not and knows not he knows not: he is a fool - shun him;
he who knows not and knows he knows not: he is simple - teach him;
he who knows and knows not he knows: he is asleep - wake him;
he who knows and knows he knows: he is wise - follow him.
Arabian proverb

Moore [Moore 1985] described a logic called *autoepistemic logic* for modeling the beliefs of an ideally reasoning agent who reflects on his own beliefs. He was particularly interested in representing arguments of the following form: 'If I had an elder brother I would have known it. I do not know it. So I do not have an elder brother.' Such arguments are 'defeasible' or 'nonmonotonic' because the agent would have to retract his conclusion (that he has no elder brother) if he becomes aware that an older brother does in fact exist.

More formally, classical propositional modal logic (see section 1.3) is monotonic in the sense that, given a wf B, every sentence A valid in the frame determined by a set of sentences Γ will still be valid in the set of models of $\Gamma \cup \{B\}$, i.e. if $\Gamma \models A$, then $\Gamma \cup \{B\} \models A$. A system which is capable of handling nonmonotonic arguments such as the above would, however, lack this property, i.e. if the set Γ of axioms is augmented by the addition of the new axiom B, then it may be the case that the wf A which was 'entailed' by Γ would no longer be 'entailed' by $\Gamma \cup \{B\}$. The possible world semantics of modal logic provides a monotonic entailment relation, hence Moore found it necessary to replace it by a new semantics of his own invention. The essential idea is that beliefs can be divided into two classes: objective beliefs concerned with the state of affairs 'out there' and subjective beliefs concerned with the internal state of affairs revealed by introspective reflection. An autoepistemic interpretation would therefore have two components, one representing the state of affairs 'outside' the agent and the other representing the internal state of affairs revealed by introspection.

2.1 Autoepistemic Interpretations

We are interested in sets of autoepistemic wfs which represent the total beliefs of an agent capable of reasoning about his own beliefs. Autoepistemic wfs are just wfs of a classical propositional modal language, but the adjective 'autoepistemic' is intended to signal that truth values are allocated to wfs in a new way. In classical propositional modal logic the

truth value of a wf A without any modal operator depends only on the assignment of truth values to atoms; the truth value of a wf $\Box A$ is determined with the help of an accessibility relation on the set of possible worlds but is (eventually) reducible to the truth of nonmodal wfs. In autoepistemic logic, however, there is no systematic connection between the truth of a wf A (as determined by an assignment of truth values to atoms) and the truth of a wf $\Box A$. The truth of a belief A where A does not contain the modal operator can (still) be checked by looking 'outwards', i.e. at the specified world (truth assignment), but the truth of a belief $\Box A$ should be checked by looking 'inward', i.e. the truth value of $\Box A$ depends on whether A is indeed believed or not. To capture this intuition $\Box A$ is evaluated with respect to a set of wfs representing the beliefs of the agent. Such a set should therefore be explicitly included in any autoepistemic interpretation.

Following Lukaszewicz [Lukaszewicz 1990], we define an *autoepistemic interpretation* of a language as a pair $\langle f, \mathcal{T} \rangle$ where f is an assignment of truth values to the atoms of the language and $\mathcal{T}$ is a set of wfs. Intuitively f specifies what is true in the outside world and $\mathcal{T}$ specifies the wfs which the agent believes.

We write $\langle f, \mathcal{T} \rangle \mid\!\vdash A$ with the intended meaning of 'The wf A is true in the interpretation $\langle f, \mathcal{T} \rangle$'. Following Moore, truth values are given to the wfs of the language by the rules below:

$$
\begin{array}{lll}
\langle f, \mathcal{T} \rangle \mid\!\vdash p & \text{iff} & f(p) = \mathrm{T} \text{ where } p \in \mathcal{P} \\
\langle f, \mathcal{T} \rangle \mid\!\vdash \neg B & \text{iff} & \langle f, \mathcal{T} \rangle \not\mid\!\vdash B \\
\langle f, \mathcal{T} \rangle \mid\!\vdash (B \vee C) & \text{iff} & \langle f, \mathcal{T} \rangle \mid\!\vdash B \text{ or } \langle f, \mathcal{T} \rangle \mid\!\vdash C \\
\langle f, \mathcal{T} \rangle \mid\!\vdash (B \wedge C) & \text{iff} & \langle f, \mathcal{T} \rangle \mid\!\vdash B \text{ and } \langle f, \mathcal{T} \rangle \mid\!\vdash C \\
\langle f, \mathcal{T} \rangle \mid\!\vdash (B \rightarrow C) & \text{iff} & \langle f, \mathcal{T} \rangle \mid\!\vdash C \text{ or } \langle f, \mathcal{T} \rangle \not\mid\!\vdash B \\
\langle f, \mathcal{T} \rangle \mid\!\vdash (B \leftrightarrow C) & \text{iff} & \text{either } \langle f, \mathcal{T} \rangle \mid\!\vdash B \text{ and also } \langle f, \mathcal{T} \rangle \mid\!\vdash C, \\
 & & \text{or } \langle f, \mathcal{T} \rangle \not\mid\!\vdash B \text{ and also } \langle f, \mathcal{T} \rangle \not\mid\!\vdash C \\
\langle f, \mathcal{T} \rangle \mid\!\vdash \Box B & \text{iff} & B \in \mathcal{T}.
\end{array}
$$

We see a wf $\Box A$ is true if and only if A is in the set $\mathcal{T}$. Thus the set $\mathcal{T}$ itself completely determines the truth of all wfs of the form $\Box A$, independently of the assignment of truth values to the atoms. So the agent and the world together determine an autoepistemic interpretation. An *autoepistemic model* of a set $\mathcal{S}$ of autoepistemic wfs is an autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ of the language in which all the wfs of $\mathcal{S}$ are true. In particular an autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ will be a model of $\mathcal{T}$ iff all the wfs of $\mathcal{T}$ are true in it, i.e. iff all the agent's beliefs are actually true in this interpretation.

Let us look at an example. Say our agent is in the control room of a nuclear power station where for simplicity we assume he has only to watch a screen with a green and a red light. We use the language with two atoms p and q where we think of p as 'The green light is on' and q as 'The red light is on'. Suppose the agent's beliefs are represented by $\mathcal{T} = \{p, q\}$, but the actual situation is that only the green light is on, i.e. $f(p) = \mathrm{T}$ but $f(q) = \mathrm{F}$. Clearly one of the agent's beliefs (namely q) is false in the interpretation, so $\langle f, \mathcal{T} \rangle$ is not a model of his beliefs. The interpretation is, however, a model of (say) the set $\mathcal{S} = \{p, q \rightarrow p, \Box q\}$, since $\langle f, \mathcal{T} \rangle \mid\!\vdash p$ and $\langle f, \mathcal{T} \rangle \mid\!\vdash q \rightarrow p$ (because $f(p) = \mathrm{T}$) and $\langle f, \mathcal{T} \rangle \mid\!\vdash \Box q$ (because $q \in \mathcal{T}$).

With this semantics in mind, we would like to find out what beliefs an ideally reasoning agent should accept on the basis of a set Γ of initial beliefs. We assume the agent to be *ideal*, in other words we wish to ignore the limitation that resource-boundedness might impose on the agent's capacity to be aware of all the consequences of his beliefs. This means the set of beliefs should be closed under an appropriate form of entailment. In a monotonic logic the set of beliefs would be $\mathrm{Cn}(\Gamma)$ or $\mathrm{Th}(\Gamma)$ where Γ is the set of initial beliefs. What is the analogue in autoepistemic logic of the semantic or deductive closure of a set Γ? Several

other questions can also be asked. For example, in classical logic every interpretation has associated with it a theory, namely the set of all sentences true in that interpretation. Given an autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ and bearing in mind that $\mathcal{T}$ is the agent's set of beliefs, we may ask whether $\langle f, \mathcal{T} \rangle$ is a model of $\mathcal{T}$. Other questions are whether notions analogous to soundness and completeness could be formulated in autoepistemic logic.

2.2 Belief Sets

The obvious way to characterise the set of beliefs entertained by an ideally reasoning agent on the basis of a set Γ of initial beliefs would be to define a suitable semantic consequence relation $\models$ or a deductive consequence relation $\vdash$ and to describe the belief set as $\mathrm{Cn}(\Gamma)$ or $\mathrm{Th}(\Gamma)$. It is, however, not obvious what would constitute a suitable consequence relation and so Moore followed a less direct approach. In fact, Moore gives three different characterisations of such belief sets: as *autoepistemic extensions*, as *stable extensions* and as the theory determined by a certain kind of possible world model (*explicit complete S5-models*).

The definition of an autoepistemic extension is primarily semantic in origin, and relies on a clever reformulation of the familiar notions of soundness and completeness. The notion of soundness in classical logic, i.e. if $\Gamma \vdash A$, then $\Gamma \models A$, or equivalently $\mathrm{Th}(\Gamma) \subseteq \mathrm{Cn}(\Gamma)$, may be formulated as follows. If we take any $A \in \mathrm{Th}(\Gamma)$, then $\Gamma \models A$, i.e. A is true in every model of Γ. In other words $\mathrm{Th}(\Gamma)$ is sound if all its members are true in every model of Γ, or equivalently, the deductive closure of Γ, $\mathrm{Th}(\Gamma)$, is sound iff every model of Γ is a model of $\mathrm{Th}(\Gamma)$. Now we are able to define soundness in autoepistemic logic. We do it with respect to a set of premises:

- A set $\mathcal{T}$ is *sound with respect to a set* Γ of premises iff every autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ which is an autoepistemic model of Γ is also an autoepistemic model of $\mathcal{T}$.

Since we are trying to model the beliefs of a rational agent it seems reasonable to require that his set of beliefs should be sound with respect to his initial beliefs: the beliefs should be true if the premises are true.

The notion of completeness in classical logic, i.e. if $\Gamma \models A$, then $\Gamma \vdash A$, or equivalently $\mathrm{Cn}(\Gamma) \subseteq \mathrm{Th}(\Gamma)$, may be formulated as follows. Any wf $A \in \mathrm{Cn}(\Gamma)$ should belong to the deductive closure $\mathrm{Th}(\Gamma)$ of Γ. Thus we require that $\mathrm{Th}(\Gamma)$ contains all wfs that are true in every model of Γ. Now we are ready to define completeness of a set:

- A set $\mathcal{T}$ is *semantically complete* iff $\mathcal{T}$ contains every wf that is true in every autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ which is an autoepistemic model of $\mathcal{T}$.

Since we assume that the agent is an ideal reasoner his set of beliefs should be semantically complete: his set of beliefs should contain all wfs that are entailed by his original beliefs and his awareness that he believes them. By this we implicitly assume that time and memory are unbounded.

From the above it seems reasonable to expect the belief set of an agent to be sound with respect to the premises, semantically complete and, finally, to include all the premises. This brings us to the definition of an autoepistemic extension of a set:

- A set $\mathcal{T}$ is an *autoepistemic extension of a set* Γ iff

  $\mathcal{T}$ is sound with respect to Γ,

  $\mathcal{T}$ is semantically complete and

  $\Gamma \subseteq \mathcal{T}$.

So, one very natural way to describe the belief set of an ideal reasoner is to regard the terms 'belief set' and 'autoepistemic extension of Γ' as synonymous.

The above approach to characterise belief sets does have the drawback of appearing to ignore the ideal reasoner's capability to perform introspection. The notion of stable sets makes this capability explicit. Informally the belief set of a rational and ideally reasoning agent should include every wf A that he could infer either by classical logic or by reflecting on what he believes and does not believe. To catch this intuition Stalnaker [Stalnaker 1980] has suggested that the belief set of an ideally rational agent should be stable where stability is defined as follows:

- A set $\mathcal{T}$ of wfs is *stable* iff $\mathcal{T}$ satisfies the following three conditions:

  if $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \vdash B$, then $B \in \mathcal{T}$,

  if $A \in \mathcal{T}$, then $\Box A \in \mathcal{T}$ and

  if $A \notin \mathcal{T}$, then $\neg\Box A \in \mathcal{T}$.

However, the problem remains of how inference should be understood. What is the meaning of $\vdash$ in the first property that stable sets should have? One straightforward approach is to interpret $\vdash$ as the semantic consequence relation involving arbitrary autoepistemic interpretations, i.e. to regard $A \vdash B$ as an abbreviation of 'B is true in every autoepistemic interpretation $\langle f, \mathcal{T}\rangle$ that is a model of A'. Since $\mathcal{T}$ varies over all sets of wfs, one is essentially looking at all possible ways to associate truth values with wfs of the form $\Box C$. An equivalent approach which has the virtue of emphasising that the truth of $\Box C$ does not in any way depend on the truth of C, is to define a new semantics for the language based on the idea of a nonmodal interpretation.

Let f be any assignment of truth values to the atoms and to the wfs of the form $\Box A$ of the language (in other words, think of wfs of the form $\Box A$ as being additional atoms). The assignment f can be extended in the usual way to a valuation v_f assigning a truth value to every wf of the language:

$$
\begin{array}{lll}
v_f(p) & = f(p) & \text{for all } p \in \mathcal{P}, \\
v_f(\Box A) & = f(\Box A) & \text{for every wf beginning with } \Box, \\
v_f(\neg A) & = \mathrm{T} & \text{iff } v_f(A) = \mathrm{F}, \\
v_f(A \vee B) & = \mathrm{T} & \text{iff } v_f(A) = \mathrm{T} \text{ or } v_f(B) = \mathrm{T}, \\
v_f(A \wedge B) & = \mathrm{T} & \text{iff } v_f(A) = \mathrm{T} \text{ and } v_f(B) = \mathrm{T}, \\
v_f(A \rightarrow B) & = \mathrm{T} & \text{iff } v_f(A) = \mathrm{F} \text{ or } v_f(B) = \mathrm{T}, \\
v_f(A \leftrightarrow B) & = \mathrm{T} & \text{iff either } v_f(A) = \mathrm{T} \text{ and } v_f(B) = \mathrm{T}, \\
 & & \text{or } v_f(A) = \mathrm{F} \text{ and } v_f(B) = \mathrm{F}.
\end{array}
$$

We call such a valuation a *nonmodal interpretation* of the language. A *nonmodal model* of a set $\mathcal{T}$ of wfs is a nonmodal interpretation of the language in which all the wfs of $\mathcal{T}$ are true. We see nonmodal interpretations and models of a set of autoepistemic wfs are precisely those we would get in ordinary propositional logic if we treat all wfs of the form

$\Box A$ as atoms. As before, a wf A is entailed by a set Γ iff A is true in every model of Γ and we write it as $\Gamma \models_{nm} A$ where the nm serves to remind us that the interpretations underlying the entailment relation are nonmodal interpretations. We inherit the soundness and completeness theorems of propositional logic, i.e. given a set Γ of wfs, $\Gamma \models_{nm} A$ iff $\Gamma \vdash A$. This means then that the first condition that must be satisfied for stability, namely 'if $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \vdash B$, then $B \in \mathcal{T}$', may be phrased equivalently as

if $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \models_{nm} B$, then $B \in \mathcal{T}$.

Whenever a wf A is entailed in this way by a set Γ of autoepistemic wfs, i.e. whenever it is the case that $\Gamma \models_{nm} A$ or (equivalently) $\Gamma \vdash A$, we will say that A *is a tautological consequence of* Γ.

Given a nonmodal interpretation, we define the associated autoepistemic interpretation corresponding with it as follows: For every a nonmodal interpretation v_f the *associated autoepistemic interpretation* is $\langle g, \mathcal{T} \rangle$ where g is the restriction of f to the set of atoms and $\mathcal{T}$ is the set of all wfs A such that $\Box A$ is true with respect to v_f. Similarly, for every autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ the *associated nonmodal interpretation* is the valuation v_g where g is defined by $g(p) = \mathrm{T}$ iff $f(p) = \mathrm{T}$, for all $p \in \mathcal{P}$, and $g(\Box A) = \mathrm{T}$ iff $A \in \mathcal{T}$, for all wfs of the form $\Box A$.

Lemma 2.1 *For every wf A, A is true in an autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ iff A is true in the associated nonmodal interpretation v_g.*

The result follows directly from the definition of associated interpretations. ♠

Let $\Gamma \models_{ae} A$ abbreviate the assertion that the set Γ autoepistemically entails the wf A, i.e. that A is true in every autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ which is an autoepistemic model of Γ. Then the following corollary follows from lemma 2.1:

Corollary 2.1 $\Gamma \models_{nm} A$ *iff* $\Gamma \models_{ae} A$, *in other words* $\models_{nm} = \models_{ae}$.

Assume $\Gamma \models_{nm} A$. From lemma 2.1 we know that all the members of the set Γ and the wf A will be true in all the associated autoepistemic interpretations of the nonmodal models of these wfs. Can it be the case that there exists some other autoepistemic interpretation in which all members of Γ are true but in which the wf A is false? No, because then the members of Γ would be true and the wf A false in the associated nonmodal interpretation and this contradicts the assumption that Γ (nonmodally) entails A. A similar argument is used to prove the corollary in the opposite direction. ♠

From the corollary above it follows that the first condition that must be satisfied for stability, namely if $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \models_{nm} B$, then $B \in \mathcal{T}$, may be phrased equivalently as

if $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \models_{ae} B$, then $B \in \mathcal{T}$.

It seems clear that the belief set of an ideal reasoner should be stable. A stable belief set contains all the beliefs of the agent in the sense that no further conclusions can be drawn. However, the stability conditions say nothing about what should be excluded from $\mathcal{T}$. If we specify merely that the belief set should be stable, beliefs that in no way can be built up from the initial beliefs of the agent are not excluded. We need a constraint specifying that the only wfs which may be members of the belief set should be the premises and those wfs required by the stability conditions. This is where groundedness in a set of premises comes in:

- A set $\mathcal{T}$ is *grounded in a set* Γ iff $\mathcal{T}$ is included in the tautological consequences of $\Gamma \cup \{\Box A \mid A \in \mathcal{T}\} \cup \{\neg\Box A \mid A \notin \mathcal{T}\}$.

From the above it seems reasonable to expect the belief set of an agent to be stable and grounded in his set of initial beliefs Γ. This brings us to the definition of a stable extension of a set:

- A set $\mathcal{T}$ is a *stable extension of a set* Γ iff

  $\Gamma \subseteq \mathcal{T}$,

  $\mathcal{T}$ is stable, and

  $\mathcal{T}$ is grounded in Γ.

So, a second very natural way to describe the belief set of an ideal reasoner is to regard the terms 'belief set' and 'stable extension of Γ' as synonymous.

Are the above-mentioned two natural ways to think of a belief set really different? In the next section we show that the two conjectures are in fact equivalent. In section 2.5 we examine the third approach which turns out to be very useful in practical applications.

2.3 Equivalence of Two Approaches

Let us assume the agent's belief set is stable. The first condition for stability ensures that the agent is able to perform all inferences, the second that he is aware of all his beliefs and the third that he is aware of everything that he does not believe. Suppose the belief set is, in addition to being stable, also consistent. Then the set will also have the following two properties:

if $\Box A \in \mathcal{T}$, then $A \in \mathcal{T}$ and

if $\neg\Box A \in \mathcal{T}$, then $A \notin \mathcal{T}$.

Why? Well, if $\Box A \in \mathcal{T}$ but $A \notin \mathcal{T}$, then $\neg\Box A \in \mathcal{T}$ (because of stability) and the set $\mathcal{T}$ would be inconsistent. Similarly, if $\neg\Box A \in \mathcal{T}$, and $A \in \mathcal{T}$, then $\Box A \in \mathcal{T}$ (because of stability) and again $\mathcal{T}$ would be inconsistent.

Suppose we have a stable and consistent belief set $\mathcal{T}$ and a nonmodal model of it. Then, by lemma 2.1, the associated autoepistemic interpretation will also be a model of $\mathcal{T}$. This fact will be needed in some of the remaining proofs of the chapter.

Now we are going to show that the truth of any wf of a stable belief set $\mathcal{T}$ in an autoepistemic interpretation $\langle f, \mathcal{T}\rangle$ depends only on the truth of the wfs without any modal operator (the so-called *objective* wfs of $\mathcal{T}$). This means that if the objective wfs of the (stable) belief set are true, all the wfs of the set are true. We need the following lemma [Lukaszewicz 1990] in the proof of the theorem:

Lemma 2.2 *For any wf A there is a wf*

$A' = C_1 \wedge C_2 \wedge \ldots \wedge C_n$

where for each C_i there exist k and r such that C_i can be written in the form

$A_1^i \vee \Box A_2^i \vee \Box A_3^i \vee \ldots \vee \Box A_k^i \vee \neg\Box A_{k+1}^i \vee \neg\Box A_{k+2}^i \vee \ldots \vee \neg\Box A_{k+r}^i$,

with A_1^i objective, which is such that A and A' have the same modal depth (i.e. their modal operators are nested to the same depth), $\vdash A \leftrightarrow A'$ and, for any set $\mathcal{T}$ closed under tautological consequence, $A \in \mathcal{T}$ iff $A' \in \mathcal{T}$.

By taking all occurrences of the form $\Box B$ as atoms, the result follows directly from the fact that every classical propositional wf is equivalent to a wf of the form $C_1 \wedge C_2 \wedge \ldots \wedge C_n$

where all $C_i, i = 1, \ldots, n$, are disjunctions of atoms or negations of atoms. ♠

Now we are ready to prove the following theorem.

Theorem 2.1 *Given a stable set $\mathcal{T}$ and an autoepistemic interpretation $\langle f, \mathcal{T}\rangle$ of the language, if the interpretation is a model of the objective wfs of $\mathcal{T}$, then it is a model of $\mathcal{T}$.*

Suppose first that $\mathcal{T}$ is a stable belief set and $\langle f, \mathcal{T}\rangle$ is an autoepistemic interpretation of the language and suppose further all the objective wfs of $\mathcal{T}$ are true in this interpretation (thus the interpretation is a model of the objective wfs). $\mathcal{T}$ is consistent, because otherwise it would include all wfs of the language but not all the objective wfs of the language are true in the interpretation (for example the wf $(p \wedge \neg p)$ is not true). Let A be any wf in $\mathcal{T}$. We may assume that A is in the form $C_1 \wedge C_2 \wedge \ldots \wedge C_n$, where, for each of these wfs C_i, there exist k and r such that the C_i can be written in the form

$A_1^i \vee \Box A_2^i \vee \Box A_3^i \vee \ldots \vee \Box A_k^i \vee \neg\Box A_{k+1}^i \vee \neg\Box A_{k+2}^i \vee \ldots \vee \neg\Box A_{k+r}^i$,

where A_1^i is an objective wf. Now we have two possibilities:

(i) At least one of $\Box A_2^i, \Box A_3^i, \ldots, \Box A_k^i, \neg\Box A_{k+1}^i, \neg\Box A_{k+2}^i, \ldots, \neg\Box A_{k+r}^i$ is in the set $\mathcal{T}$. But we know that if a wf of the form $\Box B$ is in $\mathcal{T}$, then $B \in \mathcal{T}$ (because of stability and consistency), so $\Box B$ is true in any autoepistemic interpretation $\langle f, \mathcal{T}\rangle$ of the language. Further, if a wf of the form $\neg\Box B$ is in $\mathcal{T}$, then $B \notin \mathcal{T}$ (again because of stability and consistency), so $\Box B$ is false in any autoepistemic interpretation, therefore $\neg\Box B$ is true in any autoepistemic interpretation $\langle f, \mathcal{T}\rangle$. It follows then that C_i is also true in the interpretation.

(ii) Say (i) does not hold. Because $\mathcal{T}$ is stable, it will contain all of $\neg\Box A_2^i$, $\neg\Box A_3^i$, ... , $\neg\Box A_k^i$, $\Box A_{k+1}^i$, $\Box A_{k+2}^i$, ..., $\Box A_{k+r}^i$. Then A_1^i must be a member of $\mathcal{T}$ because $C_i \in \mathcal{T}$. But A_1^i is true in the interpretation (it is an objective wf), so it follows that C_i is also true in the interpretation.

We have shown that every C_i is true in the interpretation. We may thus conclude that A is true in the interpretation and, because A was an arbitrarily chosen wf of $\mathcal{T}$, the interpretation is an autoepistemic model of $\mathcal{T}$. ♠

This theorem means then that any stable set is sound with respect to its objective wfs. Now we can prove that the (syntactic) property of stability is equivalent to the (semantic) property of completeness:

Theorem 2.2 *A set $\mathcal{T}$ of autoepistemic wfs is semantically complete iff $\mathcal{T}$ is stable.*

First we assume that we have a stable belief set $\mathcal{T}$ and want to show that this set is semantically complete, i.e. that the set contains all wfs that are true in all autoepistemic models $\langle g, \mathcal{T}\rangle$ of $\mathcal{T}$. We do it by showing that if a wf $A \notin \mathcal{T}$, then there is an autoepistemic model $\langle g, \mathcal{T}\rangle$ of $\mathcal{T}$ in which A is false. So, let A be any wf not in $\mathcal{T}$ and again we assume A is in the form $C_1 \wedge C_2 \wedge \ldots \wedge C_n$ where each C_i is in the form

$A_1^i \vee \Box A_2^i \vee \Box A_3^i \vee \ldots \vee \Box A_k^i \vee \neg\Box A_{k+1}^i \vee \neg\Box A_{k+2}^i \vee \ldots \vee \neg\Box A_{k+r}^i$,

where A_1^i is an objective wf. Since $\mathcal{T}$ is stable and $A \notin \mathcal{T}$, at least one of the C_i is not in $\mathcal{T}$. Let $C_t \notin \mathcal{T}$. This means that none of the wfs A_1^t, $\Box A_2^t$, $\Box A_3^t$, ..., $\Box A_k^t$, $\neg\Box A_{k+1}^t$, $\neg\Box A_{k+2}^t$, ..., $\neg\Box A_{k+r}^t$ can be in the set $\mathcal{T}$. We are now going to show that there exists an autoepistemic model $\langle g, \mathcal{T}\rangle$ of $\mathcal{T}$ in which all these $k + r$ wfs are false. We look first at the wf A_1^t and then at the other disjuncts.

Consider A_1^t. Because $\mathcal{T}$ is stable and $A_1^t \notin \mathcal{T}$, A_1^t cannot be a tautological consequence of the objective wfs of the set. By the completeness theorem for propositional logic, there must be a truth assignment f to the atoms of the language which is such that A_1^t is false, but

all the objective wfs of $\mathcal{T}$ are true. Now consider the autoepistemic interpretation $\langle f, \mathcal{T}\rangle$. We will have $\langle f, \mathcal{T}\rangle \not\models A_1^t$ and $\langle f, \mathcal{T}\rangle \models B$ for all objective wfs $B \in \mathcal{T}$. So, by theorem 2.1, $\langle f, \mathcal{T}\rangle$ is an autoepistemic model of $\mathcal{T}$ in which A_1^t is false.

Now consider the other disjuncts of C_t. We have already seen that none of them is in $\mathcal{T}$, thus they must all be false in any autoepistemic interpretation $\langle g, \mathcal{T}\rangle$ of $\mathcal{T}$, also in $\langle f, \mathcal{T}\rangle$. Therefore all the disjuncts of C_t are false in this autoepistemic model of $\mathcal{T}$, which means then that also C_t is false in this interpretation. We may conclude that the autoepistemic interpretation $\langle f, \mathcal{T}\rangle$ is a model of $\mathcal{T}$ in which the wf A is false.

To prove the theorem in the opposite direction, we assume $\mathcal{T}$ is semantically complete: any wf A which is true in every autoepistemic model $\langle g, \mathcal{T}\rangle$ of $\mathcal{T}$, will be in $\mathcal{T}$. Let $\langle f, \mathcal{T}\rangle$ be an arbitrary autoepistemic model of $\mathcal{T}$. If we show that a wf A is true in this model, it must be true in every autoepistemic model $\langle g, \mathcal{T}\rangle$ of $\mathcal{T}$ because the model was arbitrarily chosen. The wf A will then be in $\mathcal{T}$ because of the semantic completeness of the set. We now show that the three conditions for stability hold.

Suppose $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \models_{nm} B$. Because $\langle f, \mathcal{T}\rangle$ is an autoepistemic model of $\mathcal{T}$, $A_1, A_2, \ldots, A_n$ will all be true in the model and, because B is a tautological consequence of $A_1, A_2, \ldots, A_n$, B will also be true in the model. Thus $B \in \mathcal{T}$. Let us look at the second property that must hold and suppose $A \in \mathcal{T}$. Because $\langle f, \mathcal{T}\rangle$ is an autoepistemic interpretation, $\Box A$ will be true in it, thus $\Box A \in \mathcal{T}$. Thirdly, suppose $A \notin \mathcal{T}$. Because $\langle f, \mathcal{T}\rangle$ is an autoepistemic interpretation, $\Box A$ will be false in it, which means $\neg\Box A$ will be true, thus $\neg\Box A \in \mathcal{T}$. We therefore conclude that $\mathcal{T}$ is stable. ♠

We have seen that stability of an agent's belief set is equivalent to its semantic completeness. Now we show that groundedness is equivalent to soundness.

Theorem 2.3 *A set $\mathcal{T}$ of autoepistemic wfs is sound with respect to a set of premises Γ iff $\mathcal{T}$ is grounded in Γ.*

Suppose first that $\mathcal{T}$ is grounded in Γ. This means that every wf $A \in \mathcal{T}$ is included amongst the tautological consequences of $\Gamma \cup \{\Box B \mid B \in \mathcal{T}\} \cup \{\neg\Box B \mid B \notin \mathcal{T}\}$. We want to show that $\mathcal{T}$ is sound with respect to Γ, i.e. that every autoepistemic interpretation $\langle g, \mathcal{T}\rangle$ which is an autoepistemic model of Γ is an autoepistemic model of $\mathcal{T}$.

Suppose $\langle f, \mathcal{T}\rangle$ is any autoepistemic interpretation of the language in which all the wfs of Γ are true and suppose $A \in \mathcal{T}$. If $A \in \Gamma$, then A is trivially true in the interpretation. If A is of the form $\Box B$ where $B \in \mathcal{T}$ or if A is of the form $\neg\Box B$ where $B \notin \mathcal{T}$, then A is true in $\langle f, \mathcal{T}\rangle$ (from the definition of an autoepistemic interpretation). So, all members of $\Gamma \cup \{\Box B \mid B \in \mathcal{T}\} \cup \{\neg\Box B \mid B \notin \mathcal{T}\}$ are true in the interpretation, so all their tautological consequences are also true in the interpretation. Since all wfs of $\mathcal{T}$ are included in this set, the interpretation is an autoepistemic model of $\mathcal{T}$. But $\langle f, \mathcal{T}\rangle$ was an arbitrarily chosen interpretation in which all the wfs of Γ were true, so every autoepistemic interpretation $\langle g, \mathcal{T}\rangle$ which is an autoepistemic model of Γ is an autoepistemic model of $\mathcal{T}$, i.e. $\mathcal{T}$ is sound with respect to Γ.

To prove the theorem in the opposite direction, we assume $\mathcal{T}$ is sound with respect to Γ, i.e. every autoepistemic interpretation $\langle g, \mathcal{T}\rangle$ which is a model of Γ is an autoepistemic model of $\mathcal{T}$. We must show that $\mathcal{T}$ is grounded in Γ, i.e. that every wf $A \in \mathcal{T}$ is a tautological consequence of the set $\mathcal{T}' = \Gamma \cup \{\Box A \mid A \in \mathcal{T}\} \cup \{\neg\Box A \mid A \notin \mathcal{T}\}$.

Let v_f be any nonmodal model of $\mathcal{T}'$. Further, let g be the restriction of f to the atoms of the language. For any wf $A \in \mathcal{T}$, we will have $\Box A \in \mathcal{T}'$, so $\Box A$ will be true in v_f. Also, for any wf $A \notin \mathcal{T}$, we will have $\neg\Box A \in \mathcal{T}'$, so $\Box A$ will not be true in v_f. This means $\Box A$

is true in v_f iff $A \in \mathcal{T}$, thus $\langle g, \mathcal{T} \rangle$ is the associated autoepistemic interpretation of v_f. Because $\Gamma \subseteq \mathcal{T}'$ it follows from lemma 2.1 that $\langle g, \mathcal{T} \rangle$ is a model of Γ. So, by soundness, the interpretation is an autoepistemic model of $\mathcal{T}$. But, again from lemma 2.1, we know that every wf A that is true in $\langle g, \mathcal{T} \rangle$ is true in v_f. Because v_f was chosen arbitrarily, every wf in $\mathcal{T}$ is true in every nonmodal model of $\mathcal{T}'$. By the completeness theorem for propositional logic, every wf in $\mathcal{T}$ is therefore a tautological consequence of $\mathcal{T}'$. So, $\mathcal{T}$ is grounded in Γ. ♠

In this section we have shown that the two sets of properties that intuitively seemed appropriate for characterising the belief set of an ideal reasoner are equivalent. The belief set $\mathcal{T}$ of such an agent should be the autoepistemic extension of the set Γ of initial beliefs. These sets will grounded in Γ and will be stable, i.e. will be stable extensions of Γ.

Will it always be possible to construct a belief set from a given set of initial beliefs? The answer is no. Some sets of premises do not have any stable extensions, some have one and some have more than one stable extension. An example of a set with no stable extension is the following: Suppose the set of premises of the agent is $\Gamma = \{\neg\Box p \rightarrow p\}$ and suppose $\mathcal{S}$ is a stable set which includes the set Γ. Now $p \in \mathcal{S}$ because if $p \notin \mathcal{S}$, then $\neg\Box p \in \mathcal{S}$ which leads to a contradiction. But, if $p \in \mathcal{S}$, the set $\mathcal{S}$ would not be grounded in Γ. Why not? Well, $\mathcal{S}$ would not be sound with respect to Γ: one model of Γ is an interpretation making both p and $\neg\Box p$ false and this interpretation will certainly not be a model of $\mathcal{S}$. Thus Γ has no stable extension.

Suppose we have two agents, both ideal reasoners. The belief set of each contains objective wfs, representing beliefs about the outside world, and modal wfs. Suppose the two agents have exactly the same beliefs about the outside world. Will their belief sets be the same?

2.4 Objective Wfs Determine Stable Sets

We have seen in theorem 2.1 that any stable set is sound with respect to its objective wfs. We can go even further: a stable set is uniquely determined by its objective wfs. This is formally stated by the following theorem.

Theorem 2.4 *If two stable sets contain the same objective wfs, then the two sets are identical.*

Suppose two stable belief sets $\mathcal{T}_1$ and $\mathcal{T}_2$ contain the same objective wfs. If $\mathcal{T}_1$ contains all wfs, the objective wfs p and $\neg p$ will both be present, but then $\mathcal{T}_2$ will also consist of all wfs, so in this case we have that $\mathcal{T}_1 = \mathcal{T}_2$. So we assume the two sets are both consistent. Suppose now $A \in \mathcal{T}_1$. We want to show that then $A \in \mathcal{T}_2$. We do it by induction on the modal depth of A.

If the modal depth of A is 0, the result follows immediately because A is an objective wf. Now suppose the modal depth of A is $d > 0$ and that, if two stable belief sets contain the same objective wfs, then they contain exactly the same wfs of modal depth less than d.

We may assume that A is in the form $C_1 \wedge C_2 \wedge \ldots \wedge C_n$, where, for each of these wfs C_i, there exist k and r such that the C_i can be written in the form

$A_1^i \vee \Box A_2^i \vee \Box A_3^i \vee \ldots \vee \Box A_k^i \vee \neg\Box A_{k+1}^i \vee \neg\Box A_{k+2}^i \vee \ldots \vee \neg\Box A_{k+r}^i$,

where A_1^i is an objective wf and where none of these wfs has a modal depth of more than d. We have to consider three possible cases for each C_i:

(i) $\Box A_j^i \in \mathcal{T}_1$ for some j, $2 \leq j \leq k$. Then, because of stability and consistency, $A_j^i \in \mathcal{T}_1$. Since the modal depth of A_j^i is one less than the modal depth of $\Box A_j^i$, it is certainly less than d and so, because of the induction hypothesis, $A_j^i \in \mathcal{T}_2$. But then $\Box A_j^i \in \mathcal{T}_2$ (stability), hence also $C_i \in \mathcal{T}_2$ as it is a tautological consequence of $\Box A_j^i$.

(ii) $\neg\Box A_j^i \in \mathcal{T}_1$ for some j, $k+1 \leq j \leq k+r$. Then, because of stability and consistency, $A_j^i \notin \mathcal{T}_1$. Since the modal depth of A_j^i is one less than the modal depth of $\neg\Box A_j^i$, it is certainly less than d and so, because of the induction hypothesis, $A_j^i \notin \mathcal{T}_2$. But then $\neg\Box A_j^i \in \mathcal{T}_2$ (stability), hence also $C_i \in \mathcal{T}_2$ as it is a tautological consequence of $\neg\Box A_j^i$.

(iii) Suppose neither (i) nor (ii) holds. Because $\mathcal{T}_1$ is stable it must contain all of $\neg\Box A_2^i$, $\neg\Box A_3^i$, ..., $\neg\Box A_k^i$, $\Box A_{k+1}^i$, $\Box A_{k+2}^i$, ... $\Box A_{k+r}^i$. But A_1^i is a tautological consequence of these wfs and C_i, so that $A_1^i \in \mathcal{T}_1$. Since A_1^i is an objective wf, also $A_1^i \in \mathcal{T}_2$. But C_i is a tautological consequence of A_1^i, so we conclude that $C_i \in \mathcal{T}_2$.

In each of the three cases we have shown that $C_i \in \mathcal{T}_2$. This will be true for all $i, 1 \leq i \leq n$, so we have that $A \in \mathcal{T}_2$. Since A was chosen arbitrarily, every wf in $\mathcal{T}_1$ is also in $\mathcal{T}_2$. In the same way we can show that every wf in $\mathcal{T}_2$ is also in $\mathcal{T}_1$ from which we conclude that the two sets contain exactly the same wfs. ♠

Does this theorem mean that two agents with Γ_1 and Γ_2 as their respective sets of initial beliefs will have exactly the same belief set if Γ_1 and Γ_2 contain the same objective wfs? No. Say, for example, $\Gamma_1 = \{p, \Box\neg q\}$ and $\Gamma_2 = \{p, \Box q\}$. Then the stable extension of Γ_1 will include the wf $\neg q$ and the stable extension of Γ_2 will include the wf q, i.e. the first agent will believe $\neg q$ and the other one will believe q. According to the theorem the two agents will have identical beliefs, though, if *all* their objective beliefs are the same. So if the two sets of initial beliefs are such that (i) they contain the same objective wfs and (ii) the objective wfs which can be deduced from these sets are the same, then it will be the case that the respective belief sets are identical.

2.5 A Link with Possible World Semantics

In section 1.3.2 we described the possible world semantics of a classical propositional modal language. A model $\mathcal{M}$ was defined as a triple $\langle \mathcal{W}, \mathcal{R}, v \rangle$ for some set $\mathcal{W}$ of possible worlds, some accessibility relation $\mathcal{R} \subseteq \mathcal{W} \times \mathcal{W}$ and some valuation v specifying, for each $w \in \mathcal{W}$, which atoms are true. Let $\mathcal{T}$ be the set of all wfs valid in a possible world model $\mathcal{M}$. An obvious question now arises: Is $\mathcal{T}$ a belief set? In general the answer is no, but Moore [Moore 1984] shows that if $\mathcal{M}$ is a possible world model of a certain kind, then $\mathcal{T}$ is very close to being a belief set.

Suppose $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ is such that $\mathcal{R} = \mathcal{W} \times \mathcal{W}$. This means that $\mathcal{R}$ is an equivalence relation and further that every world is accessible from every world. Such models are called *complete S5-models*. Our first theorem is the following:

Theorem 2.5 *A set $\mathcal{T}$ of wfs is stable iff $\mathcal{T}$ is the set of all wfs which are valid in some complete S5-model.*

Suppose $\mathcal{T}$ is the set of wfs which are true in every world of some complete S5-model, $\mathcal{M}$. We want to show that $\mathcal{T}$ is stable. The first property to be checked is that of closedness under tautological consequence, i.e. we must check that, if $A_1, A_2, \ldots, A_n \in \mathcal{T}$ and $A_1, A_2, \ldots, A_n \models_{nm} B$, then $B \in \mathcal{T}$. Let $\Gamma = \{A_1, A_2, \ldots, A_n\}$ and suppose $\mathcal{M} \mid\vdash_w \Gamma$ and $\Gamma \models_{nm} B$. We want to show that $\mathcal{M} \mid\vdash_w B$. Now let f be the assignment which is such

that, for all $p \in \mathcal{P}$, we have $f(p) = \mathrm{T}$ iff $\mathcal{M} \mid\!\vdash_w p$ and, for all wfs of the form $\Box C$, we have $f(\Box C) = \mathrm{T}$ iff $\mathcal{M} \mid\!\vdash_w \Box C$. The extended valuation v_f satisfies exactly those wfs A for which it is the case that $\mathcal{M} \mid\!\vdash_w A$. (This is easily proved by noting that the assertion holds for atoms and wfs of the form $\Box C$ from the definition of f and then by using induction.) Thus it satisfies all members of Γ, hence also B (because Γ nonmodally entails B) and so $\mathcal{M} \mid\!\vdash_w B$.

The second property to be checked to establish that $\mathcal{T}$ is stable is that $\Box A \in \mathcal{T}$ if $A \in \mathcal{T}$. Well, if $A \in \mathcal{T}$ it means that A is true in every world $w \in \mathcal{W}$ which means that $\Box A$ is true in every world $w \in \mathcal{W}$, so $\Box A \in \mathcal{T}$. The third property necessary for stability is that, if $A \notin \mathcal{T}$, then $\neg\Box A \in \mathcal{T}$. If $A \notin \mathcal{T}$ it must be the case that there is some world w_k (say) where A is not true. This means then that $\Box A$ cannot be true at any world because w_k is accessible from all worlds, so $\neg\Box A$ is true at all worlds and therefore $\neg\Box A \in \mathcal{T}$. We have thus proved that $\mathcal{T}$ is stable.

To prove the theorem in the other direction, we assume we have a stable set $\mathcal{T}$. Let $\mathcal{T}_0$ be the set of all the objective wfs of $\mathcal{T}$. We want to show that $\mathcal{T}$ consists of all wfs which are true in some complete S5-model. If $\mathcal{T}$ contains all wfs of the language, these wfs are all true in the complete S5-model where $\mathcal{W}$ is empty. We thus assume $\mathcal{T}$ does not contain all wfs of the language, i.e. we assume the set is consistent. Let $\mathcal{V}$ be the set of all truth assignments f such that $f(A) = \mathrm{T}$ for all $A \in \mathcal{T}_0$. (The set $\mathcal{V}$ cannot be empty because $\mathcal{T}$ is consistent.) Suppose C is an objective, valid wf which is entailed by $\mathcal{T}_0$ in the sense of classical propositional logic, thus $\mathcal{T}_0 \models_{nm} C$. Then $\mathcal{T} \models_{nm} C$ and this means $C \in \mathcal{T}$ (because $\mathcal{T}$ is stable, it is closed under tautological consequence), thus $C \in \mathcal{T}_0$. Consider now the explicit complete S5-model $\langle \mathcal{V}, \mathcal{R}, v \rangle$ with $v(w,p) = w(p)$ for every $w \in \mathcal{V}$ and every $p \in \mathcal{P}$. (So this is the model where we identify each possible world with a particular truth assignment which will make all the objective wfs true and where v is compatible with this.) The set $\mathcal{T}_0$ is exactly the set of (objective) wfs which are true in every world of this model. By applying the first part of this theorem, this means that $\mathcal{T}_0$ contains all the objective wfs of some stable belief set $\mathcal{T}'$ (say) which are true in every world of the model. But, according to theorem 2.4, $\mathcal{T} = \mathcal{T}'$. Therefore we may conclude that $\mathcal{T}$ is the set of wfs true in every world of a complete S5-model. ♠

The theorem shows that every stable set of wfs may be characterised by an explicit complete S5-model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$, i.e. a complete S5-model in which $\mathcal{W} \subseteq \mathcal{V}$, the set of assignments of truth values to the atoms of the language, and where v is defined as $v(w,p) = w(p)$ for all $w \in \mathcal{W}$ and $p \in \mathcal{P}$. Let $\mathcal{T}$ be the set of all sentences valid in such a model $\mathcal{M}$ and let $f \in \mathcal{V}$. Then $\langle f, \mathcal{T} \rangle$ is an autoepistemic interpretation of the language. Is $\langle f, \mathcal{T} \rangle$ a model of $\mathcal{T}$? The answer is yes provided that the assignment f is chosen to be compatible with $\mathcal{M}$, i.e. provided that $f \in \mathcal{W}$. This is stated formally in the following theorem.

Theorem 2.6 *Given an explicit complete S5-model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ with $\mathcal{T}$ the stable set of wfs valid in $\mathcal{M}$, and an autoepistemic interpretation $\langle f, \mathcal{T} \rangle$. Then the following holds:*

(i) If $f \in \mathcal{W}$, then the autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ is an autoepistemic model of $\mathcal{T}$.

(ii) If the autoepistemic interpretation $\langle f, \mathcal{T} \rangle$ is an autoepistemic model of $\mathcal{T}$, then $f \in \mathcal{W}$, provided that there are only finitely many atoms in the language.

(*i*) Suppose $f \in \mathcal{W}$. (Hereby we implicitly assume the set $\mathcal{T}$ is consistent, otherwise $\mathcal{W}$

would be empty.) Then every wf $A \in \mathcal{T}$, i.e. every wf A that is valid in $\mathcal{M}$, will also be true in world f, thus A will be true in the autoepistemic interpretation $\langle f, \mathcal{T} \rangle$. For atoms, it follows directly. For more complex wfs a straightforward inductive argument suffices. By way of illustration, consider the case $A = \Box B$ and assume $A \in \mathcal{T}$. Then $B \in \mathcal{T}$ because $\mathcal{T}$ is stable and consistent. Thus A is true in $\langle f, \mathcal{T} \rangle$. This means then that the interpretation is an autoepistemic model of $\mathcal{T}$.

(*ii*) In the opposite direction, suppose $f \notin \mathcal{W}$. This means that for each world $w_i \in \mathcal{W}$, there will be some atom, p_i (say), for which $f(p_i) \neq v(w_i, p_i) = w_i(p_i)$. (The p_i are not necessarily distinct.) For each i, take either p_i (if $v(w_i, p_i) =$T) or $\neg p_i$ (if $v(w_i, p_i) =$F) and form their disjunction, B (say). This will be a finite disjunction because there are only a finite number of atoms in the language, thus only a finite number of possible worlds. The wf B will be true in every world $w \in \mathcal{W}$, so $B \in \mathcal{T}$, but it will be false in f. Thus $\langle f, \mathcal{T} \rangle$ is not an autoepistemic model of $\mathcal{T}$. ♠

We have now seen that under certain circumstances the semantics of Moore's autoepistemic logic is compatible with the possible world semantics described in chapter 1 of this dissertation. This fact gives us a useful tool for the construction of stable extensions. This will be illustrated in the next section.

2.6 Examples of Belief Sets

As mentioned before, a set of autoepistemic wfs may have none, one or more stable extensions and we looked at an example of a set with none. Let us now consider a set which has two stable extensions, using the language with two atoms, p and q. Say the premises of the agent are the set $\Gamma = \{\neg\Box p \rightarrow q, \neg\Box q \rightarrow p\}$. (Let us think of p as 'I am able to win' and of q as 'I am going to lose'. Then the agent believes that, if he does not believe that he is able to win, he will lose and, if he does not believe that he will lose, he will be able to win.) We are going to show that this set has two stable extensions, one containing the belief p (and not the belief q) and the other containing the belief q (and not the belief p). Let us examine the explicit complete S5-models that can be constructed. In only two cases will the stable set associated with the model be a stable extension of Γ.

Let $f_i, i = 1, 2, 3, 4$, be the following assignments of truth values to the atoms: $f_1(p) = f_1(q) =$T, $f_2(p) =$T and $f_2(q) =$F, $f_3(p) =$F and $f_3(q) =$T and $f_4(p) = f_4(q) =$F and, initially, $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$ where each $w_i, i = 1, 2, 3, 4$, is the world represented by f_i. The accessibility relation $\mathcal{R}$ is the cartesian product of the set of worlds, so we have an explicit complete S5-model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$. According to theorem 2.5 the set $\mathcal{T}$ of all wfs true at each of the four worlds will be a stable set. A stable extension, however, includes the premises. This is not the case with $\mathcal{T}$, for example at w_2 the wf $\neg\Box p$ is true but the wf q is false. The same problem occurs when we take any combination of three of the worlds as our set $\mathcal{W}$.

Suppose we take $\mathcal{W} = \{w_1, w_2\}$. Then $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, v \rangle$ with $\mathcal{R} = \mathcal{W} \times \mathcal{W}$ is an explicit complete S5-model. According to theorem 2.5 there will be a stable set $\mathcal{T}$ of autoepistemic wfs which are true at every $w \in \mathcal{W}$. Furthermore, the initial beliefs are both true at these two worlds and therefore belong to $\mathcal{T}$. Is the set $\mathcal{T}$ a stable extension of the premises Γ? Well, we know it is semantically complete (because it is stable) and we know the premises are included. Further, $\mathcal{T}$ is sound with respect to the premises Γ since from theorem 2.6 we know that the autoepistemic interpretations $\langle f_1, \mathcal{T} \rangle$ and $\langle f_2, \mathcal{T} \rangle$ that are models of Γ are

models of $\mathcal{T}$ because in each case $f_i \in \mathcal{W}$. So, $\mathcal{T}$ is indeed a stable extension of the initial beliefs of the agent. We see that $p \in \mathcal{T}$ because p is true at every world of the model $\mathcal{M}$ and that $q \notin \mathcal{T}$ because q is not true at every world of the model $\mathcal{M}$.

Let us write down a few of the wfs that the agent will believe in this case.

- All propositional wfs entailed by p must be included:
  $(p \vee \neg p)$, $(p \vee q)$, $(p \vee \neg q)$, p.
- Because of stability we have to include
  $\Box(p \vee \neg p)$, $\Box(p \vee q)$, $\Box(p \vee \neg q)$, $\Box p$.
- The following objective wfs are not included: $(\neg p \vee q)$, $(\neg p \vee \neg q)$, q, $(p \leftrightarrow q)$, $(p + q)$, $\neg q$, $\neg p$, $(p \wedge q)$, $(p \wedge \neg q)$, $(\neg p \wedge q)$, $(\neg \wedge \neg q)$, $(\neg p \wedge p)$, so the following wfs must be in the set because of stability:
  $\neg\Box(\neg p \vee q)$, $\neg\Box(\neg p \vee \neg q)$, $\neg\Box q$, $\neg\Box(p \leftrightarrow q)$, $\neg\Box(p + q)$, $\neg\Box\neg q$, $\neg\Box\neg p$, $\neg\Box(p \wedge q)$, $\neg\Box(p \wedge \neg q)$, $\neg\Box(\neg p \wedge q)$, $\neg\Box(\neg \wedge \neg q)$, $\neg\Box(\neg p \wedge p)$.
  (Note that $(p + q)$ abbreviates $(p \wedge \neg q) \vee (\neg p \wedge q)$.)

Suppose we take $\mathcal{W} = \{w_1, w_3\}$. Again both initial premises are true at these two worlds. $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, w \rangle$ with $\mathcal{R} = \mathcal{W} \times \mathcal{W}$ is an explicit complete S5-model and, as above, there will be a stable set $\mathcal{T}$ of autoepistemic wfs which are true at every $w \in \mathcal{W}$. The autoepistemic interpretations $\langle f_1, \mathcal{T} \rangle$ and $\langle f_3, \mathcal{T} \rangle$ are autoepistemic models of $\mathcal{T}$. The set $\mathcal{T}$ includes the premises Γ and, as in the argument above, we can show that it is a stable extension of the premises. But this set will include the wf q as a belief (because it is true at all worlds of the model $\mathcal{M}$) and not p.

We have found two different belief sets which are both stable extensions of Γ. What about other combinations of worlds: will they perhaps result in other stable extensions? If we take any other combination of two worlds, neither premise is included in the set $\mathcal{T}$ and so we do not get a stable extension. What about sets of worlds containing only one world? Let us consider $\mathcal{W} = \{w_1\}$. Both premises are included in $\mathcal{T}$, but we encounter a different problem: we have that $p \in \mathcal{T}$ and $q \in \mathcal{T}$ but the autoepistemic interpretation $\langle f_4, \mathcal{T} \rangle$ is a model of the initial premises but not of $\mathcal{T}$ since p and q are not true in the interpretation. This means then that the set is not sound with respect to the set of initial beliefs, thus $\mathcal{T}$ is not a stable extension of the premises. The same kind of problem occurs if we choose $\mathcal{W} = \{w_2\}$ or $\mathcal{W} = \{w_3\}$, and if we choose $\mathcal{W} = \{w_4\}$, the premises are not included in $\mathcal{T}$. Finally, suppose we take $\mathcal{W}$ as the empty set. The stable set $\mathcal{T}$ of wfs true at each world of this set is the set of all wfs. But although the set of all wfs does contain Γ it is not a stable extension of Γ because it is not grounded in Γ.

From the above we see that it may be the case that an ideal reasoner can end up with one of several possible belief sets, all based on the same set of premises! Moore raised the question of, given the fact that a set Γ of premises may lead to more than one stable extension, how autoepistemic logic should be viewed *as* a logic ([Moore 1985]). Suppose we were to view Γ as a set of axioms, what would be the analogue of $\mathrm{Cn}(\Gamma)$ or $\mathrm{Th}(\Gamma)$? In the case where Γ has a unique stable extension $\mathcal{T}$, it would be natural to regard $\mathcal{T}$ as the set of 'consequences' or 'theorems' of Γ. (The use of the term 'theorem' would however commit us to the construction of a suitable proof theory.) What if Γ has more than one stable extension? A proposal originally made by McDermott and Doyle in the context of their own nonmonotonic logic ([McDermott et al 1987]) would involve forming the intersection

of all the stable extensions of Γ. An outside observer, informed only of the agent's initial beliefs, would have no way of picking out the stable set that constitutes the agent's belief set. But the sentences in the intersection are precisely those that the outside observer could be sure must belong to the agent's belief set.

To illustrate the way in which this intersection captures the intuitive notion of 'consequence', albeit with the aid of a very simple example, consider formalising the argument 'If I had an elder brother I would have known it. I do not know it. So I do not have an elder brother.' The premises of the argument are taken to be the initial beliefs of the agent (in this case a single wf), so $\Gamma = \{p \rightarrow \Box p\}$ where we think of p as 'I have an elder brother'. The wf $\neg\Box p$ becomes a belief of the agent because of stability (p is not a belief). We claim that the conclusion of the argument, $\neg p$, belongs to the intersection of all stable extensions of Γ. To see this it is sufficient to construct all explicit complete S5-models and to examine their associated stable sets. Let us use the language with one atom namely p. When $\mathcal{W} = \emptyset$ the stable set of wfs true at every world of $\mathcal{W}$ (namely the set of all wfs) will not be grounded in Γ, thus will not be a stable extension. Suppose we take $\mathcal{W} = \{w_1, w_2\}$ and let $f_1(p) =$T represent the world w_1 and $f_2(p) =$F represent the world w_2, the first premise will not be valid in the model $\langle \mathcal{W}, \mathcal{R}, w \rangle$. If $\mathcal{W} = \{w_1\}$ the second premise will not be valid in the model. For $\mathcal{W} = \{w_2\}$, however, both premises will be valid in the model and we get a stable extension $\mathcal{T}$ that contains the wf $\neg p$. So the belief 'It is not the case that I have an older brother' is in the belief set of the agent. We have constructed all possible stable extensions (actually only one) and the sentence $\neg p$ is in their intersection, as claimed above.

Suppose at some stage an older brother does appear on the scene with convincing documentation. The agent, being rational, believes the evidence and revises his set Γ of premises to, for example, $\{p \rightarrow \Box p, p\}$. His belief set, namely the stable extension of Γ, will now include $p \rightarrow \Box p, p$ and $\Box p$, but will no longer include the wf $\neg p$. In this sense autoepistemic logic is nonmonotonic.

2.7 Summary

Autoepistemic logic was the first satisfactory use of modal logic to describe nonmonotonic inference, [Ginsberg 1987]. What about other features of logics of belief?

In chapter 1 we saw that providing a modal language with a possible world semantics in which the accessibility relation $\mathcal{R}$ was an equivalence relation resulted in a picture of an agent capable of positive and negative introspection, aware of all the consequences of his beliefs, and whose beliefs were all true (i.e. embodied knowledge rather than mere conviction). The notion of a stable extension in Moore's autoepistemic logic fits the picture of an agent still capable of positive and negative introspection and still aware of the consequences of his beliefs (although the word 'consequence' means something a little different). But, while in the context of a possible world semantics, sentences of the form $\Box A \rightarrow A$ were globally true, in autoepistemic logic we have a weaker version. The agent's beliefs are not necessarily true (consider, for example, the belief set containing all sentences), but the agent believes that they are: every stable extension will contain all wfs of the form $\Box A \rightarrow A$, representing the agent's confidence in himself - he believes that whenever he believes a sentence A, then A is the case.

A second difference is that autoepistemic logic provides no proof theory representing a way in which an agent might construct his belief set from a set Γ of initial beliefs. In the

context of possible worlds in chapter 1 an agent in a world w who believed a wf would be able to deduce further beliefs with the aid of valid sentences such as $\Box A \rightarrow A$ and rules of inference such as modus ponens and necessitation. This aspect is addressed in the next chapter where the approach of Levesque [Levesque 1990] is discussed.

Although Moore focusses on rational, ideally reasoning agents the semantics itself does not depend on that. For example no connection is made between the truth of a wf $\Box(p \wedge q)$ in some interpretation and the truth of the wfs $\Box p$ and $\Box q$. But how the belief set of an irrational agent or of an agent who is not an ideal reasoner should be constructed is not obvious. The sets will neither be stable nor grounded in the initial premises. Moore himself [Moore 1984] suggests that the only way out is simply to list all wfs of the form $\Box A$ which are true in some interpretation. We will return to this in a later chapter when Konolige's approach is discussed [Konolige 1986].

Chapter 3

Levesque's Version

'Tis with our judgments as our watches; none
Are just alike, yet each believes his own.
Pope.

In the previous chapter we saw that, in Moore's autoepistemic logic, the beliefs of an agent were the members of a set of wfs satisfying certain requirements or constraints, for example positive introspection. These constraints were given in the metalanguage. Is it possible to express such requirements in the object language, i.e. the relevant modal language? Let us consider the somewhat simpler constraint expressing the idea that A is believed whenever A is true. A natural first attempt would be to include all instances of the schema $A \rightarrow \Box A$ in the belief set of the agent. But a member of the belief set of the form $A \rightarrow \Box A$ actually expresses the notion 'The agent believes that if A is true, then A will be believed' instead of the constraint 'If A is true then A is believed'. Perhaps it would be easier to express constraints if it were possible to distinguish beliefs from other wfs in a syntactic way instead of semantically by having some set $\mathcal{T}$ to which they must belong. This is a key feature of Levesque's approach: if $\mathcal{T}$ is a belief set then the fact that $A \in \mathcal{T}$ is equivalent to the condition that $\Box A$ is true in the situation under consideration ([Levesque 1990]).The constraint 'A is believed whenever it is true' now seems to be adequately captured by the axiom schema $A \rightarrow \Box A$ which acts to exclude situations or worlds in which A is true but not believed.

This brings us to the second key feature of Levesque's approach, namely a generalisation of the possible world semantics. We will find a direct analog of the original idea that A is believed in a state w iff A is true in all worlds 'accessible from' w or 'similar to' w. This return to a semantics closely related to the possible world semantics produces a monotonic logic. And this is the third key feature of Levesque's approach: nonmonotonic patterns of reasoning are represented with the help of a new modal operator O, standing for 'The agent knows only ...'.

Finally, whereas Moore used a propositional language, Levesque uses a predicate language. Since it proves to be possible to capture notions like stability in Levesque's version one can view his approach as a generalisation of Moore's autoepistemic logic to predicate languages. We will first consider the propositional part and later have a brief look at the predicate part.

3.1 Belief Sets

We use the alphabet of classical modal logic with the addition of one letter namely O. In other words our alphabet is $\mathcal{S} = \mathcal{P} \cup \{\neg, \vee, \wedge, \rightarrow, \leftrightarrow, \Box, O, (,)\}$ where $\mathcal{P}$ is the set of atoms. The set $\mathcal{F}$ of wfs (sentences) of the language is defined as before with the addition of one rule, namely

if $A \in \mathcal{F}$, then $O(A) \in \mathcal{F}$

and we omit parentheses if no ambiguity can arise.

Whereas the operator $\Box$ is read as 'The agent believes that ...', the operator O is supposed to be read informally as 'The agent believes *only* that ...'. The challenge is to define a semantics so that the wf OA will be interpreted as 'The agent believes only that A' or, equivalently, 'A is all that is believed by the agent'.

The *objective* wfs of the language are those containing neither $\Box$ nor O, the *basic* wfs are those that do not contain the O operator and the *subjective* wfs are those wfs all of whose atoms occur inside the scope of $\Box$ or O. Examples of subjective wfs are $\neg\Box(p \wedge q)$, $O(q \vee \Box p)$ and $Op \rightarrow \Box p$. A wf like $p \vee \Box q$ is basic and a wf like $p \vee Oq$ is neither objective, nor basic, nor subjective.

Let w be an assignment, i.e. a function from the set $\mathcal{P}$ to the set {T,F} and let $\mathcal{V}$ be the set of all assignments. Levesque's semantics develops two ideas met in chapter 1: possible worlds will, as in the case of explicit models, be taken to be assignments, and beliefs will still be seen as sentences true relative to a set of worlds (intuitively those worlds the agent considers 'accessible'). Instead of using an accessibility relation to associate with each world w a set of similar worlds, we will work directly with pairs $\mathcal{W}, w$ in which $\mathcal{W} \subseteq \mathcal{V}$. Following Halpern and Lakemeyer [Halpern et al 1995] we call a pair $\mathcal{W}, w$ a situation and write $\mathcal{W}, w \mid\!\vdash A$ with the intended meaning of 'The situation $\mathcal{W}, w$ satisfies the wf A' or, equivalently, 'A is true in the situation $\mathcal{W}, w$'. (Note that it is not necessarily the case that $w \in \mathcal{W}$.) Formally we have the following definition:

$\mathcal{W}, w \mid\!\vdash p$	iff	$w(p) = \mathrm{T}$ where $p \in \mathcal{P}$
$\mathcal{W}, w \mid\!\vdash \neg A$	iff	$\mathcal{W}, w \not\mid\!\vdash A$
$\mathcal{W}, w \mid\!\vdash (A \vee B)$	iff	$\mathcal{W}, w \mid\!\vdash A$ or $\mathcal{W}, w \mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash (A \wedge B)$	iff	$\mathcal{W}, w \mid\!\vdash A$ and $\mathcal{W}, w \mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash (A \rightarrow B)$	iff	$\mathcal{W}, w \not\mid\!\vdash A$ or $\mathcal{W}, w \mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash A \leftrightarrow B$	iff	either $\mathcal{W}, w \mid\!\vdash A$ and also $\mathcal{W}, w \mid\!\vdash B$, or $\mathcal{W}, w \not\mid\!\vdash A$ and also $\mathcal{W}, w \not\mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash \Box A$	iff	for every $w' \in \mathcal{W}$, $\mathcal{W}, w' \mid\!\vdash A$
$\mathcal{W}, w \mid\!\vdash OA$	iff	$\mathcal{W}, w \mid\!\vdash \Box A$ and for every w', if $\mathcal{W}, w' \mid\!\vdash A$ then $w' \in \mathcal{W}$.

We see, for example, that the wf $\Box p$ is true in a situation $\mathcal{W}, w$ iff p is true at every world in $\mathcal{W}$ and the wf Op is true in a situation $\mathcal{W}, w$ iff p is true at every world in $\mathcal{W}$ *and at no other world.* The truth of a subjective wf A in a given situation $\mathcal{W}, w$ does not depend on w, so in that case we often write $\mathcal{W} \mid\!\vdash A$ instead of $\mathcal{W}, w \mid\!\vdash A$ and similarly we often write $w \mid\!\vdash A$ when A is an objective wf. Note that the truth of OA is defined in terms of $\Box A$. As soon as we have specified which wfs of the form $\Box A$, with A basic, are true, the truth or falsity of all wfs of the form OA is determined.

Informally the beliefs entertained by the agent in a situation are those sentences that are true in all worlds $w \in \mathcal{W}$. The formal definition of a belief set is the following:

- A set $\mathcal{T}$ is a *belief set for* $\mathcal{W}$ iff $\mathcal{T} = \{A \mid A \text{ is basic and } \mathcal{W} \mid\!\vdash \Box A\}$.

The key idea of the semantics above is that sentences can be matched with sets of worlds. In a language with a finite number of atomic sentences every set of worlds can be distinguished from every other set of worlds by looking at sentences true in the respective sets. However, in a language with infinitely many atomic sentences, there may be several sets of worlds in which precisely the same sentences are true. (This becomes obvious when we reflect upon cardinalities; the language has countably many sentences but there are uncountably many assignments.) Levesque thus defines the notion of equivalence between sets of worlds and gives a standard way to pick representatives of the equivalence classes. The formal definition of equivalent sets of assignments is the following:

- A set $\mathcal{W}_1$ is *equivalent to* a set $\mathcal{W}_2$ iff
  for every basic A we have that $\mathcal{W}_1 \mid\!\vdash \Box A$ iff $\mathcal{W}_2 \mid\!\vdash \Box A$.

Each set $\mathcal{W}$ of assignments has a unique largest superset $\mathcal{W}^+$ which is equivalent to it (proven below) and this set is used as the representative of the relevant equivalence class. A corollary tells us how this maximal set can be constructed from $\mathcal{W}$, namely

- $\mathcal{W}^+ = \{w \mid \text{for every basic } A, \text{ if } \mathcal{W} \mid\!\vdash \Box A, \text{ then } \mathcal{W}, w \mid\!\vdash A\}$.

Suppose $\mathcal{T}$ is a belief set for $\mathcal{W}$. Then $\mathcal{W}^+$ satisfies $\Box A$ for every $A \in \mathcal{T}$ and is maximal with respect to this property.

In the next section we prove the existence and uniqueness of a largest equivalent superset of $\mathcal{W}$.

3.2 Maximal Sets

In this section we show that the equivalence class of a set $\mathcal{W}$ of assignments always contains a unique maximal element and then illustrate the importance of these maximal elements for a suitable definition of satisfiability and entailment.

We start by defining the set $\mathcal{W}^+$, given any set $\mathcal{W}$ of assignments.

- For any set $\mathcal{W}$ of assignments, $\mathcal{W}^+ = \{w \mid \mathcal{W} \text{ is equivalent to } \mathcal{W} \cup \{w\}\}$.

From the definition it follows that the set $\mathcal{W}^+$ is the set of all assignments that can be added to the set $\mathcal{W}$ without changing any beliefs. Clearly $\mathcal{W} \subseteq \mathcal{W}^+$. It remains to show that $\mathcal{W}^+$ is equivalent to $\mathcal{W}$, that $\mathcal{W}^+$ is maximal in this regard and that the equivalence class of $\mathcal{W}$ has no other maximal members.

Lemma 3.1 *Given two equivalent sets $\mathcal{W}_1$ and $\mathcal{W}_2$, the following holds for any basic A and any world w:*
$\mathcal{W}_1, w \mid\!\vdash A$ *iff* $\mathcal{W}_2, w \mid\!\vdash A$.

The result follows immediately for atomic sentences and by induction for negations, conjunctions, et cetera, as well as for sentences of the form $\Box B$ (from the definition of equivalent sets). ♠

The second lemma is non-intuitive but the result is needed in later proofs.

Lemma 3.2 *Let $\mathcal{W}$ and $\mathcal{W}^*$ be any two sets of assignments. Suppose that for every $w \in \mathcal{W}^*$ and every basic A such that $\mathcal{W} \models \Box A$, we have that $\mathcal{W}, w \models A$. Then $\mathcal{W}$ is equivalent to $\mathcal{W}^* \cup \mathcal{W}$.*

Let $\mathcal{W}$ and $\mathcal{W}^*$ be any two sets of assignments and assume for every $w \in \mathcal{W}^*$ and every basic A such that $\mathcal{W} \models \Box A$, we have that $\mathcal{W}, w \models A$. We are going to use induction to show that for any w and any basic A,

$\mathcal{W}, w \models A$ iff $(\mathcal{W}^* \cup \mathcal{W}), w \models A$.

This clearly holds for atomic sentences and by induction for negations, conjunctions, et cetera. What about sentences of the form $\Box B$? Well, suppose firstly for some w we have $\mathcal{W}, w \models \Box B$, in other words $\mathcal{W} \models \Box B$. By the definition of satisfiability it then follows that, for every $w_1 \in \mathcal{W}$, we have $\mathcal{W}, w_1 \models B$ and, by the induction hypothesis, $(\mathcal{W}^* \cup \mathcal{W}), w_1 \models B$. But, from the initial assumption we know that because $\mathcal{W} \models \Box B$, it is the case that $\mathcal{W}, w_2 \models B$ for all $w_2 \in \mathcal{W}^*$ and hence, again by the induction hypothesis, $(\mathcal{W}^* \cup \mathcal{W}), w_2 \models B$. Thus, for every $w_3 \in \mathcal{W}^* \cup \mathcal{W}$, we have that $(\mathcal{W}^* \cup \mathcal{W}), w_3 \models B$, thus that $(\mathcal{W}^* \cup \mathcal{W}) \models \Box B$, so $(\mathcal{W}^* \cup \mathcal{W}), w \models \Box B$.

To prove the assertion in the opposite direction, suppose for some w, $(\mathcal{W}^* \cup \mathcal{W}), w \models \Box B$, in other words $(\mathcal{W}^* \cup \mathcal{W}) \models \Box B$. Then, for every $w_1 \in \mathcal{W}^* \cup \mathcal{W}$, we have $(\mathcal{W}^* \cup \mathcal{W}), w_1 \models B$, thus for every $w_2 \in \mathcal{W}$, $(\mathcal{W}^* \cup \mathcal{W}), w_2 \models B$. By induction it follows that, for every $w_2 \in \mathcal{W}$, $\mathcal{W}, w_2 \models B$, hence $\mathcal{W} \models \Box B$, so $\mathcal{W}, w \models \Box B$. ♠

Now we are ready for the theorem stating that every set of assignments has one and only one equivalent largest superset.

Theorem 3.1 *$\mathcal{W}^+$ is the unique largest superset of a set $\mathcal{W}$ of assignments which is equivalent to it.*

We first prove that $\mathcal{W}$ is equivalent to $\mathcal{W}^+$. Consider any $w \in \mathcal{W}^+$ and any basic A such that $\mathcal{W} \models \Box A$. From the definition of $\mathcal{W}^+$ it follows that $\mathcal{W}$ is equivalent to $\mathcal{W} \cup \{w\}$ and thus $(\mathcal{W} \cup \{w\}) \models \Box A$ and hence $(\mathcal{W} \cup \{w\}), w \models A$. But by lemma 3.1 then also $\mathcal{W}, w \models A$. We arbitrarily chose the world $w \in \mathcal{W}^+$ and the basic sentence A, so by lemma 3.2 it follows that $\mathcal{W}$ and $\mathcal{W} \cup \mathcal{W}^+$ are equivalent. However, $\mathcal{W} \subseteq \mathcal{W}^+$ and thus $\mathcal{W}$ is equivalent to $\mathcal{W}^+$.

Now we prove that any other set which is equivalent to $\mathcal{W}$ is a subset of $\mathcal{W}^+$. Suppose $\mathcal{W}'$ is such a set and $w \in \mathcal{W}'$. Consider any basic A such that $\mathcal{W} \models \Box A$. Then also $\mathcal{W}' \models \Box A$ because $\mathcal{W}$ and $\mathcal{W}'$ are equivalent sets. So $\mathcal{W}', w \models A$ and, by lemma 3.1, $\mathcal{W}, w \models A$. Then, by lemma 3.2, it follows that $\mathcal{W}$ is equivalent to $\mathcal{W} \cup \{w\}$. From the definition of $\mathcal{W}^+$ it then follows that $w \in \mathcal{W}^+$. So we have that $\mathcal{W}' \subseteq \mathcal{W}^+$.

We have thus succeeded in showing that $\mathcal{W}^+$ is the unique largest equivalent superset of $\mathcal{W}$. ♠

Theorem 3.1 shows that there is one and only one way to extend a set of assignments to make it as large as possible. This extension will not change the truth value of any basic sentence, so will also not change in a given situation whether a sentence is a belief or not (by lemma 3.1, since the two sets are equivalent). Such maximal sets have the following characterisation:

Corollary 3.1 *For every set $\mathcal{W}$ of assignments,*

$\mathcal{W}^+ = \{w \mid$ *for every basic A, if $\mathcal{W} \models \Box A$, then $\mathcal{W}, w \models A$*$\}$.

Let $\mathcal{W}^* = \{w \mid$ for every basic A, if $\mathcal{W} \models \Box A$, then $\mathcal{W}, w \models A\}$. We first show that

$\mathcal{W}^* \subseteq \mathcal{W}^+$. Let w be any element of $\mathcal{W}^*$. Then, for every basic A, if $\mathcal{W} \mid\!\vdash \Box A$, then $\mathcal{W}, w \mid\!\vdash A$. But from lemma 3.2 we know that then $\mathcal{W}$ is equivalent to $\mathcal{W} \cup \{w\}$, thus from the definition of $\mathcal{W}^+$ we have $w \in \mathcal{W}^+$.

Now we show that $\mathcal{W}^+ \subseteq \mathcal{W}^*$. Let w be any element of $\mathcal{W}^+$. So $\mathcal{W}$ is equivalent to $\mathcal{W} \cup \{w\}$. Consider any basic A such that $\mathcal{W} \mid\!\vdash \Box A$. Then $(\mathcal{W} \cup \{w\}) \mid\!\vdash \Box A$, therefore $(\mathcal{W} \cup \{w\}), w \mid\!\vdash A$, thus (by lemma 3.1) $\mathcal{W}, w \mid\!\vdash A$. This means that for every basic A, if $\mathcal{W} \mid\!\vdash \Box A$, then $\mathcal{W}, w \mid\!\vdash A$, and hence $w \in \mathcal{W}^*$. ♠

A set Γ is *satisfiable* iff there is a maximal set $\mathcal{W}$ and an assignment w such that, for every $A \in \Gamma$, it is the case that $\mathcal{W}, w \mid\!\vdash A$. We write $\mathcal{W}, w \mid\!\vdash \Gamma$. Further, a set Γ *entails* a wf A, written as $\Gamma \models A$, iff the set $\Gamma \cup \{\neg A\}$ is not satisfiable, and a wf A is *valid* iff it is entailed by the empty set.

Is the maximality of $\mathcal{W}$ in the definition of entailment necessary? Suppose we were to change the given definition of satisfiability by dropping the requirement that $\mathcal{W}$ be maximal. This would make no difference to basic wfs. It will, however, make a real difference in the case of a wf containing the O operator. Let us illustrate this, using a language with countably many atoms $p_i, i = 1, 2, \ldots$. Consider the wf p_1 and let $\mathcal{W}$ be the set of all assignments that map p_1 onto T. The set $\mathcal{W}$ is maximal, because if we add any world $w \notin \mathcal{W}$ to it, it will no longer be the case that $\mathcal{W} \mid\!\vdash \Box p_1$, i.e. the wf p_1 will no longer be believed. Now define the set Γ to be

$\Gamma = \{\Box A \mid A \text{ is basic and } \mathcal{W} \mid\!\vdash \Box A\} \cup \{\neg \Box A \mid A \text{ is basic and } \mathcal{W} \mid\!\vdash \neg \Box A\}$.

The notation $\mathcal{W} \mid\!\vdash \neg \Box A$ is a convenient alternative to $\mathcal{W} \not\mid\!\vdash \Box A$.

We have that $\mathcal{W} \mid\!\vdash \Gamma$. Because $\mathcal{W} \mid\!\vdash \Box p_1$ and also, if $w \mid\!\vdash p_1$ then $w \in \mathcal{W}$, it follows that $\mathcal{W} \mid\!\vdash Op_1$. Thus both $\Box p_1$ and Op_1 are satisfied by $\mathcal{W}$. Further, $\Box p_1$ is entailed by the set Γ. Is Op_1 also entailed by Γ? Yes, because $\Gamma \cup \{\neg Op_1\}$ is unsatisfiable. We show it as follows:

Suppose it is not the case, i.e. suppose $\Gamma \cup \{\neg Op_1\}$ is satisfiable. Then there is a maximal set $\mathcal{W}'$ (say) and an assignment w such that $\mathcal{W}', w \mid\!\vdash A$ for every $A \in \Gamma$ and $\mathcal{W}', w \mid\!\vdash \neg Op_1$. This means that we will have

$\mathcal{W}', w \mid\!\vdash \Box p_1$ (because $\Box p_1 \in \Gamma$) and $\mathcal{W}', w \mid\!\vdash \neg Op_1$,

i.e.

$\mathcal{W}', w \mid\!\vdash \Box p_1$ and $\mathcal{W}', w \not\mid\!\vdash Op_1$,

i.e.

$\mathcal{W}', w' \mid\!\vdash p_1$ for all $w' \in \mathcal{W}'$ and $\mathcal{W}', w'' \mid\!\vdash p_1$ for some $w'' \notin \mathcal{W}'$.

But we know $\mathcal{W}' \subseteq \mathcal{W}$ because $\mathcal{W}$ contains all w where p_1 is true, and, for the same reason, $w'' \in \mathcal{W}$. From corollary 3.1, however, we see that $\mathcal{W}'$ is not maximal since we can add w'' to $\mathcal{W}'$ because p_1 is true at w''. This contradicts the choice of $\mathcal{W}'$.

We have shown that Op_1 is entailed by Γ. Now we will show that if the maximality requirement in the definition is dropped, then Op_1 is no longer entailed by Γ. We will do this by constructing a non-maximal set $\mathcal{W}^*$ which is equivalent to the (maximal) set $\mathcal{W}$ above.

Let w_1 be the member of $\mathcal{W}$ which maps all $p_i, i = 1, 2, \ldots$ onto T and consider the set $\mathcal{W}^* = \mathcal{W} \setminus \{w_1\}$. This set is equivalent to $\mathcal{W}$ and although Levesque only states the fact, we will prove it here for completeness sake. We therefore want to show that, for every basic A, $\mathcal{W} \mid\!\vdash \Box A$ if and only if $\mathcal{W}^* \mid\!\vdash \Box A$. The result from left to right follows easily because $\mathcal{W}^* \subseteq \mathcal{W}$, but the other direction, i.e. if $\mathcal{W}^* \mid\!\vdash \Box A$, then $\mathcal{W} \mid\!\vdash \Box A$, is less simple to prove. First note that the sentence A can only contain a finite number of atoms which will be amongst $p_1, p_2, \ldots, p_m$ for some m. Let $\mathcal{F}_m \subseteq \mathcal{F}$ be the set of sentences generated by these

m atoms. In order to prove the result we need two lemmas.

Lemma 3.3 *Suppose C is a basic sentence of $\mathcal{F}_m$. Let v_1 and v_2 be two worlds coinciding on all the atoms in C. Then the following holds for any set $\mathcal{V}$ of worlds:*
$\mathcal{V}, v_1 \models C$ *iff* $\mathcal{V}, v_2 \models C$.

We are going to prove the result by induction. Suppose $\mathcal{X}$ is the subset of $\mathcal{F}_m$ containing all the basic sentences A for which the result holds. First we have to show that $\{p_1, p_2, \ldots, p_m\} \subseteq \mathcal{X}$. So, suppose $\mathcal{V}, v_1 \models p_i$ for any $1 \leq i \leq m$. Then $\mathcal{V}, v_2 \models p_i$ trivially. Similarly, if $\mathcal{V}, v_2 \models p_i$, then $\mathcal{V}, v_1 \models p_i$. Thus $p_i \in \mathcal{X}$ for every i. Suppose now $A \in \mathcal{X}$. Is it the case that $\neg A \in \mathcal{X}$? Well, assume $\mathcal{V}, v_1 \models \neg A$. Then $\mathcal{V}, v_1 \not\models A$ and by the induction hypothesis, $\mathcal{V}, v_2 \not\models A$, so $\mathcal{V}, v_2 \models \neg A$. Also, if $\mathcal{V}, v_2 \models \neg A$, then $\mathcal{V}, v_2 \not\models A$ and by the induction hypothesis, $\mathcal{V}, v_1 \not\models A$, so $\mathcal{V}, v_1 \models \neg A$. Thus $\neg A \in \mathcal{X}$. The results for the other connectives can be proved in a similar way. Let us look at the case where $A \in \mathcal{X}$ and investigate whether $\Box A \in \mathcal{X}$. Suppose $\mathcal{V}, v_1 \models \Box A$. Then we immediately get $\mathcal{V}, v_2 \models \Box A$ since v_1 and v_2 do not play any role. Similarly, if $\mathcal{V}, v_1 \not\models \Box A$, then $\mathcal{V}, v' \not\models A$ for some world $v' \in \mathcal{V}$ and we will have $\mathcal{V}, v_2 \not\models \Box A$ because (again) v_1 and v_2 do not play any role. Thus $\Box A \in \mathcal{X}$. ♠

Lemma 3.4 *Suppose C is a basic sentence of $\mathcal{F}_m$. Then the following holds with the sets $\mathcal{W}$ and $\mathcal{W}^*$ as defined above and v arbitrary:*
$\mathcal{W}, v \models C$ *iff* $\mathcal{W}^*, v \models C$.

We are going to prove the result by induction. Let $\mathcal{X}$ be the subset of $\mathcal{F}_m$ containing all the basic sentences A for which the result holds, i.e. $\mathcal{W}, v \models A$ iff $\mathcal{W}^*, v \models A$ for all worlds v. First we have to show that $\{p_1, p_2, \ldots, p_m\} \subseteq \mathcal{X}$. So suppose for any $1 \leq i \leq m$ we have $\mathcal{W}, v \models p_i$. Then $\mathcal{W}^*, v \models p_i$ and similarly in the other direction. Thus $p_i \in \mathcal{X}$ for all $i \leq m$. Suppose $A \in \mathcal{X}$. Is it the case that $\neg A \in \mathcal{X}$? Well, assume $\mathcal{W}, v \models \neg A$. Then $\mathcal{W}, v \not\models A$, i.e. $\mathcal{W}^*, v \not\models A$ (by the induction hypothesis), thus $\mathcal{W}^*, v \models \neg A$. Similarly, if $\mathcal{W}, v \not\models \neg A$, then $\mathcal{W}, v \models A$, so by the induction hypothesis $\mathcal{W}^*, v \models A$. Thus we have $\mathcal{W}^*, v \not\models \neg A$. So $\neg A \in \mathcal{X}$. Suppose $A, B \in \mathcal{X}$. Is it the case that $A \vee B \in \mathcal{X}$? Well, suppose $\mathcal{W}, v \models A \vee B$. Then $\mathcal{W}, v \models A$ or $\mathcal{W}, v \models B$, thus $\mathcal{W}^*, v \models A$ or $\mathcal{W}^*, v \models B$ by the induction hypothesis, so $\mathcal{W}^*, v \models A \vee B$. Similarly, if $\mathcal{W}^*, v \models A \vee B$, then $\mathcal{W}^*, v \models A$ or $\mathcal{W}^*, v \models B$, thus $\mathcal{W}, v \models A$ or $\mathcal{W}, v \models B$ by the induction hypothesis, so $\mathcal{W}, v \models A \vee B$. Thus $A \vee B \in \mathcal{X}$. The results for the other nonmodal connectives can be proved in a similar way. Let us look at the case where $A \in \mathcal{X}$ and investigate whether $\Box A \in \mathcal{X}$. Suppose $\mathcal{W}, v \models \Box A$. Then we have that $\mathcal{W}, v' \models A$ for all $v' \in \mathcal{W}$. By the induction hypothesis we have that $\mathcal{W}^*, v' \models A$ for all $v' \in \mathcal{W}$ and so, since $\mathcal{W}^* \subseteq \mathcal{W}$, that $\mathcal{W}^*, v' \models A$ for all $v' \in \mathcal{W}^*$, thus $\mathcal{W}^*, v \models \Box A$. In the other direction, suppose $\mathcal{W}, v \not\models \Box A$. Then there exists a world $v'' \in \mathcal{W}$ (say) such that $\mathcal{W}, v'' \not\models A$ and, by the induction hypothesis, $\mathcal{W}^*, v'' \not\models A$. Now we have two possibilities: either (i) A is not satisfied at some situation $\mathcal{W}^*, v''$ where v'' is a world in $\mathcal{W}^*$ or (ii) A is satisfied at $\mathcal{W}^*, w$ for all worlds $w \in \mathcal{W}^*$ and not satisfied at $\mathcal{W}^*, v''$ where $v'' \notin \mathcal{W}^*$. In the first case, i.e. if $\mathcal{W}^*, v'' \not\models A$ where $v'' \in \mathcal{W}^*$, we can conclude that $\mathcal{W}^*, v \not\models \Box A$. Suppose, however, we have case (ii). Then v'' can only be world w_1, i.e. the world mapping every atom p_i of $\mathcal{F}$ onto T. Now we construct a world namely w_2 coinciding with w_1 on all the atoms of $\mathcal{F}_m \subseteq \mathcal{F}$. So, define $w_2 \in \mathcal{W}^* \subseteq \mathcal{W}$ as the world that maps $p_1, p_2, \ldots, p_m$ onto T and all other atoms onto F. Now we have $\mathcal{W}^*, w_1 \not\models A$, but $\mathcal{W}^*, w_2 \models A$. But this is impossible according to lemma 3.3. We conclude thus that $v'' \in \mathcal{W}^*$ and so $\mathcal{W}^*, v \not\models \Box A$. Thus we have that $\Box A \in \mathcal{X}$. ♠

At last we are in a position to show that if $\mathcal{W}^* \mid\!\vdash \Box A$, then $\mathcal{W} \mid\!\vdash \Box A$. But now the equivalence of the two sets $\mathcal{W}$ and $\mathcal{W}^*$ follows immediately from lemma 3.4 above.

We know that $\mathcal{W}^*$ is *not* maximal. We still have $\mathcal{W}^* \mid\!\vdash \Gamma$ (easily seen by applying lemma 3.4) and $\Gamma \models \Box p_1$, but it will no longer be the case that $\Gamma \models Op_1$. Why not? Let us see:

$\mathcal{W}^*, w_1 \mid\!\vdash p_1$

$\mathcal{W}^* \not\mid\!\vdash Op_1$ (because there is a world which is not a member of $\mathcal{W}^*$ where p_1 is true)

$\mathcal{W}^* \mid\!\vdash \neg Op_1$

$\mathcal{W}^* \mid\!\vdash \Gamma \cup \neg Op_1$,

which means that the set $\Gamma \cup \{\neg Op_1\}$ is satisfiable.

This example shows that if, when discussing satisfiability, we do not specify that the set $\mathcal{W}$ of worlds should be maximal, we may have a set of worlds that is artificially small: worlds are excluded but there is no additional information to account for it. In the example world w_1 was excluded but there is no basic wf that is true in the smaller set of worlds and not true in the original set of worlds. By restricting ourselves to maximal sets this problem is avoided. We conclude the section with the following summarising theorem:

Theorem 3.2 *The maximality requirement in the definition of satisfiability is a necessary condition.*

The result follows immediately from the counterexample and lemmas 3.3 and 3.4 above. ♠

3.3 Stability

In order to relate a belief set as defined by Levesque to the belief sets of Moore's autoepistemic logics we need the notion of stability, which depends on a suitable entailment relation. As in chapter 2 we use valuations which treat sentences of the form $\Box A$ or OA as atoms. Levesque uses a slightly different notation where he splits the function into two disjoint sets: an assignment w mapping the members of $\mathcal{P}$ onto the set {T,F} and a second function θ mapping all wfs of the form $\Box A$ and OA onto {T,F}. Let $\theta, w \mid\!\vdash_{nm} A$ abbreviate the assertion that the pair θ, w *nonmodally satisfies* (called 'first-orderly satisfies' by Levesque) the wf A. Then the following rules hold:

$\theta, w \mid\!\vdash_{nm} p$	iff	$w(p) = T$ where $p \in \mathcal{P}$
$\theta, w \mid\!\vdash_{nm} \neg A$	iff	$\theta, w \not\mid\!\vdash_{nm} A$
$\theta, w \mid\!\vdash_{nm} (A \vee B)$	iff	$\theta, w \mid\!\vdash_{nm} A$ or $\theta, w \mid\!\vdash_{nm} B$
$\theta, w \mid\!\vdash_{nm} (A \wedge B)$	iff	$\theta, w \mid\!\vdash_{nm} A$ and $\theta, w \mid\!\vdash_{nm} B$
$\theta, w \mid\!\vdash_{nm} (A \rightarrow B)$	iff	$\theta, w \not\mid\!\vdash_{nm} A$ or $\theta, w \mid\!\vdash_{nm} B$
$\theta, w \mid\!\vdash_{nm} A \leftrightarrow B$	iff	either $\theta, w \mid\!\vdash_{nm} A$ and also $\theta, w \mid\!\vdash_{nm} B$, or $\theta, w \not\mid\!\vdash_{nm} A$ and also $\theta, w \not\mid\!\vdash_{nm} B$
$\theta, w \mid\!\vdash_{nm} \Box A$	iff	$\theta(\Box A) = \mathrm{T}$
$\theta, w \mid\!\vdash_{nm} OA$	iff	$\theta(OA) = \mathrm{T}$.

A set Γ is nonmodally satisfied by θ, w iff for every $A \in \Gamma$ we have that $\theta, w \mid\!\vdash_{nm} A$. A set Γ nonmodally entails a wf A iff the set $\Gamma \cup \{\neg A\}$ is not nonmodally satisfiable and we write it as $\Gamma \models_{nm} A$. If a set of wfs is satisfiable, it will be nonmodally satisfiable. (We simply specify θ in such a way that it is compatible with $\mathcal{W}$: $\theta(\Box A) = \mathrm{T}$ iff $\mathcal{W}, w \mid\!\vdash \Box A$ and

$\theta(OA)$ =T iff $\mathcal{W}, w \mid\!\vdash OA$.)

Now we are able to give the definition of a stable set:

- A set $\mathcal{T}$ of basic wfs is *stable* iff the following three conditions hold:
  if $\mathcal{T} \models_{nm} A$, then $A \in \mathcal{T}$,
  if $A \in \mathcal{T}$, then $\Box A \in \mathcal{T}$ and
  if $A \notin \mathcal{T}$, then $\neg\Box A \in \mathcal{T}$.

As before we see a stable set is closed under (nonmodal) entailment and positive and negative introspection.

We are going to show that a stable set corresponds exactly to the notion of a belief set as defined by Levesque but first have to prove another result for which we need the notion of an adjunct of a set.

- A set $\mathcal{S}$ is an *adjunct of a set* $\mathcal{T}$ iff
  $\mathcal{S} = \{\Box A \mid A \text{ is basic and } A \in \mathcal{T}\} \cup \{\neg\Box A \mid A \text{ is basic and } A \notin \mathcal{T}\}$.

We give some interesting and useful results connected with an adjunct of a set.

Lemma 3.5 *A stable set includes its adjunct.*
The result follows immediately from the definitions of adjunct and stability. ♠

Lemma 3.6 *Suppose $\mathcal{T}$ is a belief set of $\mathcal{W}$ and $\mathcal{S}$ the adjunct of $\mathcal{T}$. Then $\mathcal{W} \mid\!\vdash \mathcal{S}$.*
For $A = \Box B \in \mathcal{S}$ we have that $B \in \mathcal{T}$ (from the definition of an adjunct) and thus $\mathcal{W} \mid\!\vdash \Box B$, i.e. $\mathcal{W} \mid\!\vdash A$. For $A = \neg\Box B \in \mathcal{S}$ we have that $B \notin \mathcal{T}$ (from the definition of an adjunct) and thus $\mathcal{W} \not\mid\!\vdash \Box B$, thus $\mathcal{W} \mid\!\vdash \neg\Box B$, i.e. $\mathcal{W} \mid\!\vdash A$. ♠

We know that every belief set $\mathcal{T}$ is defined in terms of a set $\mathcal{W}$ of assignments and by theorem 3.1 we know that $\mathcal{W}$ can be taken as maximal without loss of generality. Is it possible that distinct belief sets can be associated with the same $\mathcal{W}$ or that distinct maximal sets can be associated with the same belief set $\mathcal{T}$? The answer is no, proved in the following lemma, suggested by corollary 2.9 in [Levesque 1990]:

Lemma 3.7 *The mapping between maximal sets of assignments and belief sets is bijective.*
From the definition of a belief set it follows directly that distinct belief sets cannot be associated with the same maximal set of assignments.

To prove the lemma in the opposite direction, let $\mathcal{T} = \{A \mid A \text{ is basic and } \mathcal{W} \mid\!\vdash \Box A\}$ with $\mathcal{W}$ maximal and with adjunct $\mathcal{S} = \{\Box A \mid A \text{ is basic and } A \in \mathcal{T}\} \cup \{\neg\Box A \mid A \text{ is basic and } A \notin \mathcal{T}\}$. Then $\mathcal{W} \mid\!\vdash \mathcal{S}$ from lemma 3.6. Suppose now some other maximal set $\mathcal{W}^*$ of assignments is such that $\mathcal{T} = \{A \mid \mathcal{W}^* \mid\!\vdash \Box A\}$. Then it will also be the case that $\mathcal{W}^* \mid\!\vdash \mathcal{S}$, again from lemma 3.6. But then $\mathcal{W}$ and $\mathcal{W}^*$ are equivalent and hence identical because both are maximal. Why are they equivalent? Well, if $\mathcal{W} \mid\!\vdash \Box B$, then $B \in \mathcal{T}$ and so $\mathcal{W}^* \mid\!\vdash \Box B$, and if $\mathcal{W}^* \mid\!\vdash \Box B$, then $B \in \mathcal{T}$ and so $\mathcal{W} \mid\!\vdash \Box B$. ♠

Lemma 3.8 *Suppose $\mathcal{S}$ is a set of basic sentences which contains an adjunct to a stable set. Then $\mathcal{S}$ is satisfiable iff it is nonmodally satisfiable.*

We know that if $\mathcal{S}$ is satisfiable, then it will be nonmodally satisfiable. To prove the theorem in the other direction, we assume $\mathcal{S}$ contains an adjunct to a stable set $\mathcal{T}$ and is nonmodally satisfiable by θ, w, so $\theta, w \mid\!\vdash_{nm} \mathcal{S}$. Let $\mathcal{W} = \{w' \mid \theta, w' \mid\!\vdash_{nm} \mathcal{T}\}$. We are going to show that for any w' and any basic wf A we have that $\mathcal{W}, w' \mid\!\vdash A$ iff $\theta, w' \mid\!\vdash_{nm} A$. It clearly holds for atomic sentences (because then only w' plays any role) and by induction also for any objective sentence. What about sentences of the form $\Box A$? Let us see.

Suppose first that $\theta, w' \mid\!\vdash_{nm} \Box A$, i.e. $\theta(\Box A) =$T. Can it be the case that $\neg\Box A \in \mathcal{S}$? No, because if it were the case we would have that $\neg\Box A$ is nonmodally satisfied by θ, w', so $\Box A$ will not be nonmodally satisfied which means that $\theta(\Box A)$ would have to be F. But this then means that $A \in \mathcal{T}$. (Why? Well, if $A \notin \mathcal{T}$, then $\neg\Box A$ would be in the adjunct of $\mathcal{T}$ which is a subset of $\mathcal{S}$.) From the definition of $\mathcal{W}$ it then follows that for every $w' \in \mathcal{W}$ we have that $\theta, w' \mid\!\vdash_{nm} A$. By induction then $\mathcal{W}, w' \mid\!\vdash A$ and so $\mathcal{W} \mid\!\vdash \Box A$.

Suppose secondly that $\theta, w' \not\mid\!\vdash_{nm} \Box A$, i.e. $\theta(\Box A) =$F. Reasoning as above we then have that $\Box A \notin \mathcal{S}$ and so $A \notin \mathcal{T}$. But $\mathcal{T}$ is a stable set and thus closed under nonmodal entailment, so $\mathcal{T} \cup \{\neg A\}$ is nonmodally satisfiable. This means that there must be a pair θ^*, w' such that $\theta^*, w' \mid\!\vdash_{nm} \mathcal{T} \cup \{\neg A\}$. But the functions θ and θ^* cannot differ on basic wfs: both satisfy all wfs in $\mathcal{T}$ and $\mathcal{T}$ contains either $\Box A$ or $\neg\Box A$ for all basic A (by stability). So we have $\theta, w' \mid\!\vdash_{nm} \mathcal{T} \cup \{\neg A\}$. But this means then that $w' \in \mathcal{W}$ (from the definition of $\mathcal{W}$). Thus there exists a $w' \in \mathcal{W}$ such that $\theta, w' \mid\!\vdash_{nm} \neg A$. By induction then $\mathcal{W}, w' \mid\!\vdash \neg A$. Therefore $\mathcal{W} \mid\!\vdash \neg\Box A$.

So, for every w' we have that $\mathcal{W}, w' \mid\!\vdash A$ iff $\theta, w' \mid\!\vdash_{nm} A$. Therefore we have $\mathcal{W}, w \mid\!\vdash \mathcal{S}$, thus $\mathcal{S}$ is satisfiable. ♠

Suppose $\mathcal{S}$ is a set of basic sentences containing an adjunct to a stable set and A is any basic sentence. It then follows from the above theorem that $\mathcal{S} \models A$ iff $\mathcal{S} \models_{nm} A$. Now we are able to prove that stable sets and belief sets (as defined by Levesque) are one and the same.

Theorem 3.3 *Suppose $\mathcal{T}$ is a set of basic sentences. Then $\mathcal{T}$ is stable iff $\mathcal{T}$ is a belief set.*

Suppose first that $\mathcal{T}$ is a belief set for $\mathcal{W}$, i.e. $\mathcal{T} = \{A \mid A \text{ is basic and } \mathcal{W} \mid\!\vdash \Box A\}$. Say $\mathcal{T} \models_{nm} A$. Is it the case that $A \in \mathcal{T}$? Well, suppose it is not the case. Then $\mathcal{W} \not\mid\!\vdash \Box A$, so there exists a $w \in \mathcal{W}$ such that $\mathcal{W}, w \not\mid\!\vdash A$ which means that $\mathcal{W}, w \mid\!\vdash \neg A$. We know $\mathcal{W}, w \mid\!\vdash \mathcal{T}$, so $\mathcal{T} \cup \{\neg A\}$ is satisfiable, thus also nonmodally satisfiable. But this means that $\mathcal{T}$ does not entail A and this is a contradiction. What about the other two properties that must be satisfied for $\mathcal{T}$ to be a stable set? Suppose $A \in \mathcal{T}$. Then $\Box A \in \mathcal{T}$, because if $\mathcal{W} \mid\!\vdash \Box A$, then also $\mathcal{W} \mid\!\vdash \Box\Box A$. Suppose $A \notin \mathcal{T}$. We want to show that $\neg\Box A \in \mathcal{T}$. Suppose this is not the case, i.e. $\mathcal{W} \not\mid\!\vdash \Box\neg\Box A$. Then there exists a $w \in \mathcal{W}$ such that $\mathcal{W}, w \not\mid\!\vdash \neg\Box A$, thus $\mathcal{W}, w \mid\!\vdash \Box A$ but this is a contradiction. So we have shown that a belief set is stable.

To prove the theorem in the other direction we assume $\mathcal{T}$ is stable. We have two cases: the set is either satisfiable or unsatisfiable. Suppose $\mathcal{T}$ is satisfiable. Then there exists a situation $\mathcal{W}, w$ such that $\mathcal{W}, w \mid\!\vdash \mathcal{T}$. For any basic wf A, if $A \in \mathcal{T}$, then it follows from stability that $\Box A \in \mathcal{T}$, so $\mathcal{W} \mid\!\vdash \Box A$. If, on the other hand, $A \notin \mathcal{T}$, then it follows from stability that $\neg\Box A \in \mathcal{T}$, so $\mathcal{W} \mid\!\vdash \neg\Box A$. Therefore $A \in \mathcal{T}$ iff $\mathcal{W} \mid\!\vdash \Box A$, in other words $\mathcal{T}$ is a belief set.

Suppose now $\mathcal{T}$ is unsatisfiable. We noted above (lemma 3.5) that the adjunct of a stable set is contained in the set, so $\mathcal{T}$ contains the adjunct to $\mathcal{T}$. Then it follows from lemma 3.8 that $\mathcal{T}$ is also not nonmodally satisfiable. This means that, for every basic sentence A, we have that $\mathcal{T} \cup \{\neg A\}$ is not nonmodally satisfiable, so $\mathcal{T} \models_{nm} A$ for every basic A. By

the definition of stability it follows that $A \in \mathcal{T}$, thus $\mathcal{T}$ contains every basic wf. So in this case $\mathcal{T}$ is the belief set of the empty set of assignments. ♠

We have now shown that belief sets as defined by Levesque are stable sets and vice versa. As in the case of Moore's autoepistemic logic the sets are uniquely determined by their objective subsets. But Levesques's belief sets are not the same as the belief sets defined by Moore. In addition to being stable Moore's belief sets were grounded in a set Γ of initial premises so that only beliefs that could be derived in some way from Γ were entertained. How does Levesque handle this? This is discussed in the next section.

3.4 Relation to Stable Extensions

We saw that one of the (equivalent) ways in which Moore describes the belief set of an agent is to specify it as the stable extension of a set Γ of initial premises. Levesque also gives a definition of a stable extension (equivalent to Moore's definition):

- A set $\mathcal{T}$ is a *stable extension of a set* Γ iff $\mathcal{T}$ satisfies the equation

  $\mathcal{T} = \{A \mid A$ is basic and

  $\Gamma \cup \{\Box B \mid B \in \mathcal{T}\} \cup \{\neg\Box B \mid B \notin \mathcal{T}\} \models_{nm} A\}$.

Suppose we are given a maximal set $\mathcal{W}$ of assignments with $\mathcal{T}$ as its belief set and $\mathcal{S}$ as the adjunct to $\mathcal{T}$ and further suppose the set Γ has only one member namely the sentence B. Then the stable extension of $\{B\}$ relative to $\mathcal{W}$ will be the set of all basic wfs which are nonmodally entailed by the union of $\{B\}$ and $\mathcal{S}$. Now we are ready to prove that a stable extension corresponds with only knowing.

Theorem 3.4 *For any basic sentence A and any maximal set $\mathcal{W}$ of assignments, $\mathcal{W} \mid\!\vdash OA$ iff the belief set of $\mathcal{W}$ is a stable extension of $\{A\}$.*

Let $\mathcal{W}$ be a given set of assignments and let $\mathcal{T}$ be its belief set and $\mathcal{S}$ the adjunct to $\mathcal{T}$. We thus have that $\mathcal{W} \mid\!\vdash \mathcal{S}$ (from lemma 3.6). We want to show that $\mathcal{W} \mid\!\vdash OA$ iff $\mathcal{T}$ is the set of all basic wfs that are nonmodally entailed by $\{A\} \cup \mathcal{S}$. To do this we may use $\models$ instead of $\models_{nm}$, from lemma 3.8. Thus we need to show the following:

$\mathcal{W} \mid\!\vdash OA$ iff for every basic B we have $B \in \mathcal{T}$ iff $\{A\} \cup \mathcal{S} \models B$,

or, equivalently,

$\mathcal{W} \mid\!\vdash OA$ iff for every basic B we have $\mathcal{W} \mid\!\vdash \Box B$ iff $\{A\} \cup \mathcal{S} \models B$.

We first assume that $\mathcal{W} \mid\!\vdash OA$. We have to prove an 'if' and an 'only if' part. For the 'if' part, assume $\{A\} \cup \mathcal{S} \models B$ and let w be any element of $\mathcal{W}$. Since $\mathcal{W} \mid\!\vdash OA$, we have that $\mathcal{W}, w \mid\!\vdash \{A\} \cup \mathcal{S}$, and therefore, $\mathcal{W}, w \mid\!\vdash B$. So, for every $w \in \mathcal{W}$ we have that $\mathcal{W}, w \mid\!\vdash B$ which means that $\mathcal{W} \mid\!\vdash \Box B$.

For the 'only if' part, assume $\mathcal{W} \mid\!\vdash \Box B$. To show that $\{A\} \cup \mathcal{S} \models B$, let $\mathcal{W}'$ be any maximal set of assignments and w be any assignment. Suppose $\mathcal{W}', w \mid\!\vdash (\{A\} \cup \mathcal{S})$. Then $\mathcal{W}' = \mathcal{W}$ because $\mathcal{S}$ is an adjunct of the belief set for $\mathcal{W}$ (so any maximal set $\mathcal{W}'$ which is such that $\mathcal{W}' \mid\!\vdash \mathcal{S}$ must be equivalent to $\mathcal{W}$ and hence equal to it). This means then that $\mathcal{W}, w \mid\!\vdash A$, thus $w \in \mathcal{W}$ because $\mathcal{W} \mid\!\vdash OA$. But if $w \in \mathcal{W}$, then $\mathcal{W}, w \mid\!\vdash B$ because $\mathcal{W} \mid\!\vdash \Box B$. So we have shown that for any $\mathcal{W}'$ and w, if $\mathcal{W}', w \mid\!\vdash (\{A\} \cup \mathcal{S})$, then $\mathcal{W}', w \mid\!\vdash B$, therefore $\{A\} \cup \mathcal{S}) \models B$.

Now assume that for every basic B we have $\mathcal{W} \mid\!\vdash \Box B$ iff $\{A\} \cup \mathcal{S} \models B$. First we must show that $\mathcal{W} \mid\!\vdash \Box A$, i.e. that A is true at every $w \in \mathcal{W}$. It follows directly from the fact that

$\{A\} \cup \mathcal{S} \models A$. Next we must show that if $\mathcal{W}, w \mid\!\vdash A$ then $w \in \mathcal{W}$. So suppose $\mathcal{W}, w \mid\!\vdash A$. Then $\mathcal{W}, w \mid\!\vdash \{A\} \cup \mathcal{S}$ because $\mathcal{W} \mid\!\vdash \mathcal{S}$. Now consider any B such that $\mathcal{W} \mid\!\vdash \Box B$. We have that $\{A\} \cup \mathcal{S} \models B$ (from our assumption) and thus $\mathcal{W}, w \mid\!\vdash B$, so $w \in \mathcal{W}^+$ (from the definition of $\mathcal{W}^+$). Because $\mathcal{W}$ is maximal, $w \in \mathcal{W}$. Thus we have shown that for every w, if $\mathcal{W}, w \mid\!\vdash A$, then $w \in \mathcal{W}$, therefore $\mathcal{W} \mid\!\vdash OA$. ♠

From this theorem then follows that 'only knowing a wf A' means that the total beliefs of an agent are the members of a stable extension of $\{A\}$, in other words what is believed is derivable from A using nonmodal logic and introspection. We have a semantic account (closely related to the semantics of possible worlds) for the (syntactic) notion of stable extensions which Moore used in describing nonmonotonic logic. It is also possible to state how many stable extensions of a wf A exist:

Theorem 3.5 *A sentence A has exactly as many stable extensions as there are maximal sets of assignments where OA is true.*

We know from lemma 3.7 that the mapping between maximal sets of assignments and belief sets is bijective. Thus the mapping between maximal sets of assignments and stable sets is bijective (from theorem 3.3) and thus (from theorem 3.4 above) there will be a stable extension of A corresponding with every maximal set of assignments. ♠

3.5 Proof Theory

By making use of the operator O derivations can be done inside the logic. In order to do this a proof theory must be given. Such a theory for the basic part of the language (i.e. wfs without the O operator) is formed by the following axiom schemas and inference rules:

(1) All substitution instances of valid sentences.
(2a) $\Box A$ where A is a valid sentence.
(3a) $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$.
(4a) $(A \rightarrow \Box A)$ where A is *subjective*.
(5) From A and $(A \rightarrow B)$, infer B.

What about O? Before giving axioms and rules involving the 'only know' operator, we add an additional letter, namely N, to the alphabet and an additional rule for constructing the set $\mathcal{F}$ of wfs, namely

if $A \in \mathcal{F}$, then $N(A) \in \mathcal{F}$.

and we omit parentheses if no ambiguity can arise. The addition of N will simplify the rules and axioms involving O. We give the following definition:

$\mathcal{W}, w \mid\!\vdash NA$ iff for every $w' \notin \mathcal{W}$ we have $\mathcal{W}, w' \mid\!\vdash A$.

Keeping in mind the definitions of $\Box A$ and OA, we are able to write O in terms of $\Box$ and N: $OA = (\Box A \wedge N\neg A)$. Let $\overline{\mathcal{W}}$ be the set of all assignments w not in $\mathcal{W}$. Then we have

$\mathcal{W}, w \mid\!\vdash NA$ iff for every $w' \in \overline{\mathcal{W}}$ we have $\mathcal{W}, w' \mid\!\vdash A$.

We see that the operators N and $\Box$ are similar with the important difference that the definition of satisfaction of a wf starting with either of them is in terms of the complement of the set of worlds where the other is satisfied. In possible world terms we range over inaccessible worlds when dealing with N and accessible worlds when dealing with $\Box$. Now we are ready for the axiom schemas and rules of inference where we include all the modal

operators. Subjective wfs are redefined as wfs where all atoms $p \in \mathcal{P}$ occur inside the scope of $\Box$ or O or N.

(1) All substitution instances of valid sentences.
(2*a*) For every valid sentence A, $\Box A$.
(2*b*) For every valid sentence A, NA.
(3*a*) $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$.
(3*b*) $N(A \rightarrow B) \rightarrow (NA \rightarrow NB)$.
(4*a*) $(A \rightarrow \Box A)$ where A is *subjective*.
(4*b*) $(A \rightarrow NA)$ where A is *subjective*.
(5) From A and $(A \rightarrow B)$, infer B.
(6) For every objective sentence A which is not valid, $NA \rightarrow \neg\Box A$.
(7) For every sentence A, OA abbreviates $(\Box A \wedge N\neg A)$.

A wf B is provable from a set Γ of wfs (written as $\Gamma \vdash B$) iff there exists a sequence $A_1, A_2, \ldots, A_n$ of n wfs such that each A_i is an instance of one of the above axiom schemas or is a member of Γ or is formed according to one of the above inference rules. Levesque proved the completeness of the proof architecture in detail and sketched the proof for soundness. Let us investigate the validity of the rules (2*b*), (3*b*) and (4*b*) above:

(2*b*) For every valid sentence A, NA:
If a sentence A is valid it means that $\mathcal{W}, w \mid\!\vdash A$ holds for every situation $\mathcal{W}, w$, thus also for all situations $\mathcal{W}, w'$ where $w' \notin \mathcal{W}$, hence for all situations $\mathcal{W}, w$ we will have $\mathcal{W}, w \mid\!\vdash NA$.

(3*b*) $N(A \rightarrow B) \rightarrow (NA \rightarrow NB)$:
Such a sentence can only be false if for some situation $\mathcal{W}, w$, we have that $\mathcal{W}, w \mid\!\vdash N(A \rightarrow B)$ and $\mathcal{W}, w \not\mid\!\vdash (NA \rightarrow NB)$. Suppose it is the case. Then $\mathcal{W}, w' \mid\!\vdash A \rightarrow B$ for all $w' \notin \mathcal{W}$ and this can only be the case if for every $w' \notin \mathcal{W}$, either $\mathcal{W}, w' \not\mid\!\vdash A$ or $\mathcal{W}, w' \mid\!\vdash B$. But we assumed $\mathcal{W}, w \not\mid\!\vdash (NA \rightarrow NB)$, i.e. $\mathcal{W}, w \mid\!\vdash NA$ and $\mathcal{W}, w \not\mid\!\vdash NB$, thus there is a $w' \notin \mathcal{W}$ such that $\mathcal{W}, w' \mid\!\vdash A$ and $\mathcal{W}, w' \not\mid\!\vdash B$. This is a contradiction, hence the original sentence cannot be false, i.e. is valid.

(4*b*) $A \rightarrow NA$ where A is subjective:
Such a sentence can only be false if for some situation $\mathcal{W}, w$, we have that $\mathcal{W}, w \mid\!\vdash A$ and $\mathcal{W}, w \not\mid\!\vdash NA$. Suppose it is the case. Then, because A is subjective, we have $\mathcal{W} \mid\!\vdash A$. Further, $\mathcal{W}, w' \not\mid\!\vdash A$ for some $w' \notin \mathcal{W}$. But again w' does not play any role, thus we have $\mathcal{W} \not\mid\!\vdash A$. This is a contradiction, hence the original sentence cannot be false, i.e. is valid.

In the next section we illustrate the usefulness of the given proof theory.

3.5.1 An Example

Let $\mathcal{K}$ be the knowledge base of an agent, in this case containing only one member namely the objective wf p. This wf is believed to be true and does not entail the wf q. Suppose the agent believes that q is false unless required to be true by what is believed. Suppose, finally, that this is all that is believed. Is it the case that q is believed to be false? In other words, if $O(p \wedge (\neg\Box q \rightarrow \neg q))$, is it the case that $\Box\neg q$?

Let us first approach the problem semantically. The argument could be as follows: Assume $\mathcal{W}$ is a maximal set of assignments such that $\mathcal{W} \mid\!\vdash O(p \wedge (\neg\Box q \rightarrow \neg q))$ and let w be some assignment such that $w \mid\!\vdash (p \wedge \neg q)$. (There exists such an assignment because p does not entail q.) Thus $\mathcal{W}, w \mid\!\vdash (p \wedge (\neg\Box q \rightarrow \neg q))$ and so $w \in \mathcal{W}$. Further, $\mathcal{W} \mid\!\vdash \neg\Box q$ because $w \mid\!\vdash \neg q$. But for every $w' \in \mathcal{W}$ we have that $\mathcal{W}, w' \mid\!\vdash (\neg\Box q \rightarrow \neg q)$, so also $w' \mid\!\vdash \neg q$. Therefore $\mathcal{W} \mid\!\vdash \Box\neg q$.

Let us tackle the same problem syntactically:

1.	$O(p \wedge (\neg\Box q \rightarrow \neg q))$	Assumption
2.	$\Box(p \wedge (\neg\Box q \rightarrow \neg q))$	Line 1, and (7)
3.	$\Box(\neg\Box q \rightarrow \neg q)$	Line 2
4.	$\Box\neg\Box q \rightarrow \Box\neg q$	Line 3, and ($3a$)
5.	$\neg\Box q \rightarrow \Box\neg\Box q$	($4a$)
6.	$\neg\Box q \rightarrow \Box\neg q$	Lines 4 and 5
7.	$N\neg(p \wedge (\neg\Box q \rightarrow \neg q))$	Line 1, and (7)
8.	$N(\neg p \vee \neg(\neg\Box q \rightarrow \neg q))$	Line 7
9.	$N(p \rightarrow \neg(\neg\Box q \rightarrow \neg q))$	Line 8
10.	$N(p \rightarrow \neg(\Box q \vee \neg q))$	Line 9
11.	$N(p \rightarrow (\neg\Box q \wedge q))$	Line 10
12.	$N(p \rightarrow q)$	Line 11
13.	$\neg\Box(p \rightarrow q)$	Line 12, and (6)
14.	$\neg(\Box p \rightarrow \Box q)$	Line 13, and ($3a$)
15.	$\Box p$	Line 2
16.	$\neg\Box q$	Lines 14 and 15, and (5)
17.	$\Box\neg q$	Lines 6 and 16, and (5)

So we have shown that it is indeed the case that the agent believes that q is not true and this was done inside the logic itself.

3.6 Predicate Logic

Let us consider a richer language than a propositional one. As before we do not include any function constants in the first order languages that we are going to consider, but the alphabet $\mathcal{S}$ does include the following symbols:

- for each positive integer n, zero, one or more predicate constants of arity n, indicated by P_i^n,
- a special two-place equality symbol, $=$,
- a countably infinite set of individual variables, indicated by x_i,
- a countably infinite set of individual constants called standard names, indicated by a_i,
- the connectives $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$,
- the modal operators $\Box, O, N$,
- punctuation symbols (and),

- quantifier symbols $\forall$ and $\exists$.

Levesque allows the occurrence of propositional letters in the alphabets of these languages but we exclude that possibility without any loss of generality. Except for the addition of the equality symbol and the three modal operators $\Box$, O and N, Levesque's languages also differ from the predicate languages of chapter 1 in that the individual constants are taken to designate members of the universe of discourse distinctly and exhaustively, i.e. interpretations are restricted to those with countably infinite universes of discourse. Without loss of generality we may assume the set of standard names to be precisely the universe of discourse. (Such interpretations are called *term models*). Wfs are formed in the obvious way and scope is defined as before. There is no restriction on the scopes of quantifiers or modal operators. A sentence is a wf with no free variables. (So, again, there are wfs that are not sentences.) The equality symbol and the standard names are considered to be logical symbols. Objective sentences are those without any $\Box$, O and N operators. Basic sentences are those without an O operator and subjective wfs are those where all predicate constants occur within the scope of a $\Box$ or O or N operator. Finally, again similar to the substitution approach in chapter 1, we write A^x_a to indicate the formula consisting of A with all free occurrences of x replaced by standard name a. Let $\mathcal{A}$ be the set of atomic sentences, i.e. predicate constants in which no variables appear.

By defining the language as above it is possible to combine the semantics met in chapter 1 with Levesque's semantics involving situations. In chapter 1 an interpretation (or world) of a predicate language was defined as a triple (ϕ, f, U) where U is the universe of discourse and ϕ a mapping from the individual constants of the language to U. We now restrict ourselves to interpretations where U is the set of standard names and ϕ is the identity. f can be any function from the set of atomic sentences $\mathcal{A}$ to the set {T,F}. Since U and ϕ are fixed, interpretations are determined by the assignment f. In accordance with our previous notation we use w instead of f.

Let w be a function of the atomic sentences to the set {T,F} and let $\mathcal{W}$ be a set of assignments. As before we write $\mathcal{W}, w \mid\!\vdash A$ to abbreviate the assertion that the sentence A is true in the situation $\mathcal{W}, w$ or (equivalently) that the situation $\mathcal{W}, w$ satisfies the sentence A. It is not necessarily the case that $w \in \mathcal{W}$. Formally we have the following definition:

$\mathcal{W}, w \mid\!\vdash A$	iff	$w(A) = T$ where $A \in \mathcal{A}$
$\mathcal{W}, w \mid\!\vdash (a_i = a_j)$	iff	a_i is the same name as a_j
$\mathcal{W}, w \mid\!\vdash \neg A$	iff	$\mathcal{W}, w \not\mid\!\vdash A$
$\mathcal{W}, w \mid\!\vdash (A \vee B)$	iff	$\mathcal{W}, w \mid\!\vdash A$ or $\mathcal{W}, w \mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash (A \wedge B)$	iff	$\mathcal{W}, w \mid\!\vdash A$ and $\mathcal{W}, w \mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash (A \rightarrow B)$	iff	$\mathcal{W}, w \not\mid\!\vdash A$ or $\mathcal{W}, w \mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash A \leftrightarrow B$	iff	either $\mathcal{W}, w \mid\!\vdash A$ and also $\mathcal{W}, w \mid\!\vdash B$, or $\mathcal{W}, w \not\mid\!\vdash A$ and also $\mathcal{W}, w \not\mid\!\vdash B$
$\mathcal{W}, w \mid\!\vdash \forall x(A)$	iff	for all a_i, $\mathcal{W}, w \mid\!\vdash (A^x_{a_i})$
$\mathcal{W}, w \mid\!\vdash \exists x(A)$	iff	for some a_i, $\mathcal{W}, w \mid\!\vdash (A^x_{a_i})$
$\mathcal{W}, w \mid\!\vdash \Box A$	iff	for every $w' \in \mathcal{W}$, $\mathcal{W}, w' \mid\!\vdash A$
$\mathcal{W}, w \mid\!\vdash NA$	iff	for every $w' \notin \mathcal{W}$, $\mathcal{W}, w' \mid\!\vdash A$
$\mathcal{W}, w \mid\!\vdash OA$	iff	$\mathcal{W}, w \mid\!\vdash \Box A$ and for every w', if $\mathcal{W}, w' \mid\!\vdash A$ then $w' \in \mathcal{W}$.

The definitions of the equivalence of sets $\mathcal{W}$ and $\mathcal{W}'$ of assignments, of satisfiability and of entailment are the same as in the propositional case. What about nonmodal satisfaction?

We only have to replace the rule involving proposition constants with a rule involving atomic sentences and add two rules involving quantifiers to those given in section 3.3, namely:

$$\begin{array}{lll} \theta, w \mid\!\vdash_{nm} A & \text{iff} & w(A) = T \text{ where } A \in \mathcal{A} \\ \theta, w \mid\!\vdash_{nm} \forall x A & \text{iff} & \mid\!\vdash_{nm} (A^x_{a_i}) \text{ for every } a_i \\ \theta, w \mid\!\vdash_{nm} \exists x A & \text{iff} & \mid\!\vdash_{nm} (A^x_{a_i}) \text{ for some } a_i. \end{array}$$

All the other definitions stay the same. The theorems about the correspondence between stable sets and belief sets and the correspondence between 'only knowing' and a stable extension which were proved in sections 3.3 and 3.4 still hold. There is, however, one major different result, namely stable sets are no longer uniquely determined by their objective subsets.

What effect does the enrichment of the language have on the proof theory given in section 3.5? First of all we have to add one rule to those listed in section 3.5:

(8) From $A^x_{a_1}, A^x_{a_2}, \ldots, A^x_{a_k}$ infer $\forall x A$ if the a_i range over all names in A and one not in A.

The new proof theory is again sound (proved by Levesque) but it is, contrary to what he expects, no longer complete. This was proved by Halpern and Lakemeyer [Halpern et al 1995]. In the next section we give an example of the use of the proof theory when we show that an agent does believe that the ever-popular Tweety flies.

3.6.1 An Example

Assume that an agent believes that Tweety is a bird and that if a bird can be consistently believed to fly, then it flies. Is it possible to show, using classical logic, that the agent believes that Tweety flies? The answer to this question is no, because at the very least the agent has to believe that Tweety can be consistently believed to fly. If, however, it is known that *all* that is believed is that Tweety is a bird, it is possible to draw the conclusion that the agent believes that Tweety flies. It is done as follows, based on an abbreviated version in [Levesque 1990].

We use the language with two predicate constants, both of arity one and indicated by *Bird* and *Fly*, and one standard name, indicated by *tweety*. Assume the agent has the knowledge base

$\mathcal{K} = \{K\} = \{Bird(tweety)\}$

and let A be the sentence

$A = \forall x((Bird(x) \wedge \neg\Box\neg Fly(x)) \rightarrow Fly(x))$.

We want to show that $O(K \wedge A) \rightarrow \Box Fly(tweety)$.

1.	$O(K \wedge A)$	Assumption
2.	$\Box(K \wedge A)$	Line 1, and (7)
3.	$\neg\Box\neg Fly(tweety) \rightarrow \Box\neg\Box\neg Fly(tweety)$	$(4a)$
4.	$\Box A$	Line 2
5.	$\Box((Bird(tweety) \wedge \neg\Box\neg Fly(tweety)) \rightarrow Fly(tweety))$	Line 4
6.	$\Box Bird(tweety)$	Line 2
7.	$\Box(\neg\Box\neg Fly(tweety) \rightarrow Fly(tweety))$	Lines 5 and 6
8.	$\Box(\neg\Box\neg Fly(tweety)) \rightarrow \Box Fly(tweety)$	Line 7, and $(3a)$
9.	$\neg\Box\neg Fly(tweety) \rightarrow \Box Fly(tweety)$	Lines 3 and 8
10.	$N\neg(K \wedge A)$	Line 1, and (7)
11.	$N(\neg K \vee \neg A)$	(1)
12.	$N(K \rightarrow \neg A)$	(1)
13.	$N(K \rightarrow \exists x\neg((Bird(x) \wedge \neg\Box\neg Fly(x)) \rightarrow Fly(x)))$	Line 12, and (1)
14.	$N(K \rightarrow \exists x\neg Fly(x))$	Line 13, and (1)
15.	$\neg\Box(K \rightarrow \exists x\neg Fly(x))$	Line 14, and (6)
16.	$\neg\Box(\neg K \vee \exists x\neg Fly(x))$	Line 15, and (1)
17.	$\neg\Box(\exists x\neg Fly(x))$	Line 16
18.	$\neg\Box(\neg Fly(tweety))$	Line 17
17.	$\Box Fly(tweety)$	Lines 9 and 18, and (5).

Note that this theorem depends on the fact that the knowledge base does not entail that Tweety does not fly, i.e. it depends on the fact that $K \rightarrow \exists x\neg Fly(x)$ is not valid. (Lines 14 - 17.) When $K = Bird(tweety)$ this is certainly true, but what will happen when the knowledge base is such that some birds may be unable to fly? Let us look at the situation where we have two birds, Tweety and Chirpy, and the agent believes that Chirpy cannot fly. Will he still believe that Tweety can fly? Yes. We show it as follows, using the same language as above but with an additional standard name, *chirpy*. The knowledge base is

$\mathcal{K} = \{Bird(tweety), Bird(chirpy), \neg Fly(chirpy)\}$.

Let

$K = Bird(tweety) \wedge Bird(chirpy) \wedge \neg Fly(chirpy)$

and

$A = \forall x((Bird(x) \wedge \neg\Box\neg Fly(x)) \rightarrow Fly(x))$.

We want to show that $O(K \wedge A) \rightarrow \Box Fly(tweety)$. The proof is again based on an abbreviated version in [Levesque 1990].

1.	$O(K \wedge A)$	Assumption
2.	$\Box(K \wedge A)$	Line 1, and (7)
3.	$\neg\Box\neg Fly(tweety) \rightarrow \Box\neg\Box\neg Fly(tweety)$	(4)
4.	$\Box A$	Line 2
5.	$\Box((Bird(tweety) \wedge \neg\Box\neg Fly(tweety)) \rightarrow Fly(tweety))$	Line 4
6.	$\Box(Bird(tweety))$	Line 2
7.	$\Box(\neg\Box\neg Fly(tweety) \rightarrow Fly(tweety))$	Lines 5 and 6
8.	$\Box(\neg\Box\neg Fly(tweety)) \rightarrow \Box Fly(tweety)$	Line 7, and (3*a*)
9.	$\neg\Box\neg Fly(tweety) \rightarrow \Box Fly(tweety)$	Lines 3 and 8
10.	$\Box\neg Fly(chirpy)$	Line 2
11.	$N\Box\neg Fly(chirpy)$	(2*b*)
12.	$N\neg(K \wedge A)$	Line 1, and (7)
13.	$N(\neg K \vee \neg A)$	Line 12 and (1)
14.	$N(K \rightarrow \neg A)$	Line 13, and (1)
15.	$N(K \rightarrow \exists x\neg((Bird(x) \wedge \neg\Box\neg Fly(x)) \rightarrow Fly(x)))$	Line 14, and (1)
16.	$N(K \rightarrow \exists x((\neg\Box\neg Fly(x)) \wedge \neg Fly(x)))$	Line 15 and (1)
17.	$N(K \rightarrow \exists x((x \neq chirpy) \wedge \neg Fly(x)))$	Lines 11 and 16, and (1)
18.	$\neg\Box(K \rightarrow \exists x((x \neq chirpy) \wedge \neg Fly(x)))$	Line 17, and (6)
19.	$\neg\Box\neg Fly(tweety)$	Line 18, and (1)
20.	$\Box Fly(tweety)$	Lines 9 and 19, and (5).

3.7 Only Know About ...

In the example above the agent argues as follows: 'If all that I know is that Tweety is a bird and a default about birds, then ...', i.e. the argument starts from the sentence

$O(Bird(tweety) \wedge \forall x((Bird(x) \wedge \neg\Box\neg Fly(x)) \rightarrow Fly(x))$

The agent's knowledge base K is the set $\{Bird(tweety)\}$. But, while $\Box(...)$ is too weak to draw the conclusion that $\Box Fly(tweety)$, $O(...)$ is too strong: we are able to arrive at the belief that Tweety flies but in addition to that it forces us to rule out all other beliefs, even it they are totally unrelated to birds and their ability (or not) to fly. It would be more convenient if the argument could start with 'If I believe the default and further all I know *about Tweety* is that he is a bird, then ...'. Levesque suggested that the operator O could be modified to identify the subject matter of the belief. Lakemeyer, [Lakemeyer 1991], investigated this further.

Returning to a propositional language with the previously defined semantics, we add an infinite set of modal operators: an operator $O\langle\Phi\rangle$ for each finite set Φ of atoms of the language. Suppose we have a situation $\mathcal{W}, w$, when will a wf $O\langle\Phi\rangle A$ be true? In other words, given $\mathcal{W}, w$, when is it the case that A is all that is believed about the atoms in Φ? In order to answer the question we first define a set of worlds $\mathcal{W}_\Phi$ which one may think of as obtained from $\mathcal{W}$ by forgetting everything that is irrelevant to Φ. Before we can formally define $\mathcal{W}_\Phi$ we need a lemma for which we need the notions of literals and clauses. *Literals* are either atoms or negated atoms and *clauses* are disjunctions of literals.

Lemma 3.9 *Belief sets are uniquely determined by their objective clauses.*

The result follows directly from the facts that belief sets are uniquely determined by the objective sentences they contain, that beliefs have equivalent conjunctive normal forms

and that $\Box(A_1 \wedge A_2 \wedge \ldots \wedge A_n) \leftrightarrow \Box A_1 \wedge \Box A_2 \wedge \ldots \wedge \Box A_n$ is valid. ♠

We want to characterise what, relative to $\mathcal{W}$, is believed about Φ. Suppose $\Phi = \{p\}$. By the above lemma we may restrict ourselves to the question of what clauses relative to $\mathcal{W}$ are believed about p. Say one of the beliefs in the belief set for $\mathcal{W}$ is a clause C which mentions p or $\neg p$, does this qualify as a belief about p? No, it is not sufficient: Suppose all the agent knows is q. Then he also believes the clause $p \vee q$, but that does not tell us anything about p. (Actually the agent knows nothing about p.) Thus for a clause C to be a belief about p, it must not depend on other beliefs which do not mention p. We go even further, namely in order to characterise all the beliefs about p we consider only minimal clauses in the sense that, while C is believed, no clause C' contained in C is believed. Formally we define such clauses as follows:

- Given a set of worlds $\mathcal{W}$ and a clause C, C is called *$\mathcal{W}$-minimal* iff
  $\mathcal{W} \mid\!\vdash \Box C$ and for all clauses C' contained in C, we have $\mathcal{W} \mid\!\vdash \Box C'$.

Let us look at an example, using the language with two atoms p and q. The possible clauses are p, $\neg p$, q, $\neg q$, $p \vee \neg p$, $p \vee q$, $p \vee \neg q$, $\neg p \vee q$, $\neg p \vee \neg q$, $q \vee \neg q$, $p \vee \neg p \vee q$, ... Let $w_1(p) = w_1(q) =$T, $w_2(p) =$T and $w_2(q) =$F, $w_3(p) =$F and $w_3(q) =$T, $w_4(p) = w_4(q) =$F. Then, for $\mathcal{W} = \{w_1\}$, the $\mathcal{W}$-minimal clauses are p and q. For $\mathcal{W} = \{w_1, w_2\}$ the $\mathcal{W}$-minimal clauses are p and $q \vee \neg q$. For $\mathcal{W} = \{w_1, w_2, w_3\}$ the $\mathcal{W}$-minimal clauses are $p \vee \neg p$ and $q \vee \neg q$ and $p \vee q$, et cetera.

- The beliefs relative to $\mathcal{W}$ about a set of atoms Φ are now defined as those $\mathcal{W}$-minimal clauses which are believed and which mention p for some $p \in \Phi$.

Now we are ready to define the set of worlds characterised by these beliefs:
Given a set of worlds $\mathcal{W}$,

- $\mathcal{W}_\Phi = \{w \mid w \mid\!\vdash C$ for all $\mathcal{W}$-minimal clauses C such that C mentions some $p \in \Phi\}$.

So the worlds in $\mathcal{W}_\Phi$ are those worlds in which everything which does not follow from the beliefs about Φ is automatically 'forgotten'. Note that it will always be the case that $\mathcal{W} \subseteq \mathcal{W}_\Phi$.

At last we are ready to give the semantics of these new operators:

- $\mathcal{W}, w \mid\!\vdash O\langle\Phi\rangle A$ iff $\mathcal{W}_\Phi, w \mid\!\vdash OA$ and $\mathcal{W} \mid\!\vdash \Box A$.

Lakemeyer shows that the new concept $O\langle\Phi\rangle$ has reasonable properties. Here are four interesting ones:

(i) The sentence $O\langle\Phi\rangle A \rightarrow \Box A$ is valid, i.e. if all the agent knows about the atoms in Φ is A, then he definitely believes A.

The result follows immediately from the definition of the satisfaction of a wf starting with $O\langle\Phi\rangle$.

(ii) Unless $\neg O\langle\Phi\rangle A$ is valid, the sentence $O\langle\Phi\rangle A \rightarrow OA$ is not valid. The reason for this is that the agent may, for example, believe something about an atom q in neither Φ nor A.

We show that the result holds by giving a counterexample. Let $A = p$, $\Phi = \{p\}$ and $\mathcal{W} = \{w \mid w \models p \wedge q\}$. The only $\mathcal{W}$-minimal clause that contains p is p itself, thus $\mathcal{W}_\Phi = \{w \mid w \models p\}$. We see $\mathcal{W} \models \Box p$ and $\mathcal{W}_\Phi \models Op$, hence $\mathcal{W} \models O(\Phi)p$. But $\mathcal{W} \not\models Op$. (Actually $\mathcal{W} \models O(p \wedge q)$.)

(iii) The sentence $O(\{p,q\})p \rightarrow O(\{q\})(q \vee \neg q)$ is valid, i.e. if all the agent knows about p and q is p, then he actually knows nothing about q.

Let $\Phi = \{p,q\}$ and $\mathcal{W} \models O(\Phi)p$. Assume $\mathcal{W} \not\models O(\{q\})(q \vee \neg q)$. Then, because $\mathcal{W} \models \Box(q \vee \neg q)$, we have that $\mathcal{W}_{\{q\}} \not\models O(q \vee \neg q)$, thus $\mathcal{W}_{\{q\}}$ is not the set of all worlds. This means that there exists a clause C which is $\mathcal{W}$-minimal and mentions q but which is not a tautology. This clause does not contain p because p is itself $\mathcal{W}$-minimal. According to the definition of $\mathcal{W}_\Phi$ we have that $\mathcal{W}_\Phi \models \Box C$. But $\mathcal{W}_\Phi = \{w \mid w \models p\}$. So we have a world $w' \in \mathcal{W}_\Phi$ which does not satisfy C (because C does not contain p) which is a contradiction.

(iv) The sentence $\neg O(\{p\})q$ is valid, i.e. something totally independent of p cannot be all the agent knows about p.

Assume that there does exist a set $\mathcal{W}$ such that $\mathcal{W} \models O(\{p\})q$. Then we have that $\mathcal{W}_{\{p\}} \models Oq$, i.e. $\mathcal{W}_{\{p\}} = \{w \mid w \models q\}$. Also, $\mathcal{W} \models \Box q$ and therefore q is $\mathcal{W}$-minimal and thus no $\mathcal{W}$-minimal clause containing p can contain q. So there exists a world w' (say) that satisfies all $\mathcal{W}$-minimal clauses mentioning p but does not satisfy q. This world must be a member of the set $\mathcal{W}_{\{p\}}$ but this is a contradiction.

We conclude this section by showing that the result about Tweety follows easily, using a propositional language where for convenience two of the atoms are indicated by *bird* and *fly*. (We think of *bird* as 'Tweety is a bird' and of *fly* as 'Tweety can fly'.) Let $\Phi = \{bird, fly\}$. Now let A be the wf

$((bird \wedge \neg\Box\neg fly) \rightarrow fly) \wedge bird.$

Let $\mathcal{W}$ be some set of worlds where this wf is believed, thus $\mathcal{W}, w \models \Box A$, i.e. A is one of the members of the belief set for $\mathcal{W}$. We want to show that

$\mathcal{W}, w \models O(\Phi)A \rightarrow \Box fly.$

Which worlds are members of the set $\mathcal{W}_\Phi$? Well, we know that if $\mathcal{W}, w \models O(\Phi)A$, then $\mathcal{W}_\Phi \models OA$, i.e. $\mathcal{W}_\Phi, w \models A$ for all $w \in \mathcal{W}_\Phi$ and for no $w \notin \mathcal{W}_\Phi$. This means that we must have $w \models bird$ for every $w \in \mathcal{W}_\Phi$. Furthermore, we must have $\mathcal{W}_\Phi, w \models (bird \wedge \neg\Box\neg fly) \rightarrow fly$ for all $w \in \mathcal{W}_\Phi$. If $w \models fly$ for every world $w \in \mathcal{W}_\Phi$, then everything works out. Suppose, however, at some world $w' \in \mathcal{W}_\Phi$ we have that $w' \not\models fly$. Is it then the case that $\mathcal{W}_\Phi, w' \not\models bird \wedge \neg\Box\neg fly$? We know *bird* is true at w', so the question is whether $\mathcal{W}_\Phi, w' \not\models \neg\Box\neg fly$. Well, if this is the case, it means that $\mathcal{W}_\Phi, w' \models \Box\neg fly$, thus that $w'' \models \neg fly$ for *all* $w'' \in \mathcal{W}_\Phi$. Thus we conclude that the set $\mathcal{W}_\Phi$ is such that either $w \models bird \wedge fly$ for every $w \in \mathcal{W}_\Phi$ or $w \models bird \wedge \neg fly$ for every $w \in \mathcal{W}_\Phi$. We assert that the first case holds. Why? Well, suppose $w \models bird \wedge \neg fly$ for every $w \in \mathcal{W}_\Phi$. Then there exists a world $w' \notin \mathcal{W}_\Phi$ such that $w' \models bird \wedge fly$ and also $\mathcal{W}_\Phi, w' \models (bird \wedge \neg\Box\neg fly) \rightarrow fly$. But this is a contradiction because A may not be true at a world outside $\mathcal{W}_\Phi$. Therefore $\mathcal{W}_\Phi = \{w \mid w \models bird \wedge fly\}$.

Now assume $\mathcal{W} \models O(\Phi)A$. Then $\mathcal{W}_\Phi \models OA$. But *fly* is true in every world of $\mathcal{W}_\Phi$ (as seen above), so $\mathcal{W}_\Phi \models \Box fly$. Since $\mathcal{W} \subseteq \mathcal{W}_\Phi$, it follows that $\mathcal{W} \models \Box fly$. We have therefore proved that, if the agent believes only A about *bird* and *tweety*, he believes that Tweety flies.

3.8 Generalisation to Arbitrary Sentences

All the theorems in this chapter have only dealt with basic sentences or sentences like OA with A a basic sentence. Levesque [Levesque 1990] generalised these theorems to deal with arbitrary sentences. We redefine the notions of belief sets and stability.

- A set $\mathcal{T}$ is a *generalised belief set for* $\mathcal{W}$ iff $\mathcal{T} = \{A \mid A \text{ is any sentence and } \mathcal{W} \mid\!\!\vdash \Box A\}$.
- A set $\mathcal{T}$ of sentences is a *generalised stable set* iff the following three conditions hold:
  if $\mathcal{T} \models A$, then $A \in \mathcal{T}$,
  if $A \in \mathcal{T}$, then $\Box A \in \mathcal{T}$ and
  if $A \notin \mathcal{T}$, then $\neg\Box A \in \mathcal{T}$.

The only change in the definition is the closure under entailment (instead of nonmodal entailment).

Theorem 3.6 *A set $\mathcal{T}$ of wfs is a generalised belief set iff $\mathcal{T}$ is a generalised stable set.*

The proof is identical to the proof of theorem 3.3 but without the diversion through lemma 3.8. ♠

- A set $\mathcal{T}$ is a *generalised stable extension of a set* Γ iff $\mathcal{T}$ satisfies the equation
  $\mathcal{T} = \{A \mid A \text{ is basic and}$
  $\Gamma \cup \{\Box B \mid B \in \mathcal{T}\} \cup \{\neg\Box B \mid B \notin \mathcal{T}\} \models A\}$.

The only difference is that the entailment relation is no longer nonmodal.

Theorem 3.7 *For any wf A and any maximal set $\mathcal{W}$ of assignments, $\mathcal{W} \mid\!\!\vdash OA$ iff the generalised belief set of $\mathcal{W}$ is a generalised stable extension of $\{A\}$.*

The proof is similar to the proof of theorem 3.4. ♠

3.9 Summary

We have seen that Levesque's version of autoepistemic logic has several advantages when compared with Moore's. It is, for example, possible to express constraints on the agent's beliefs directly in the relevant modal language. An important shortcoming is that the agent is (still) an ideal reasoner. This aspect is addressed in the next chapter when we look at the deduction structures defined by Konolige. Furthermore, the semantics as defined by Konologe makes provision for more than one agent - another issue not investigated by Levesque. (Lakemeyer [Lakemeyer 1991a], though, does address the issue of only knowing in the framework of a multi-agent autoepistemic logic. This later paper provides a formalisation of all an agent knows about a certain subject matter based on possible world semantics in such a logic.)

Chapter 4

The Deduction Model of Konolige

Chuangtse and Hueitse had strolled onto the bridge over the Hao, when the former observed, 'See how the small fish are darting about! That is the happiness of fish.' 'You are not a fish yourself,' said Hueitse. 'How can you know the happiness of the fish?' 'And you not being I,' retorted Chuangste, 'how can you know that I do not know?'
Chuangtse, c. 300 B.C.

Unlike the systems of Moore and Levesque the deduction model devised by Konolige ([Konolige 1986]) makes provision for more than one agent and in addition these agents need not be ideal reasoners. In the case of Moore and Levesque the belief sets of all agents were assumed to be closed under logical consequence: the agents were aware of all the consequences of their beliefs. But from experience we know that neither human nor robot agents are really ideal reasoners. Because of constraints on time and space there are inferences which are logically possible but which the agent does not make. Furthermore the agent's rules by which he derives 'new' beliefs are often incomplete so that even with unlimited time he cannot arrive at all the consequences.

The possible world approach is compatible with ideal reasoning. So by using such a semantics we can predict what consequences an agent can possibly derive but not what consequences will actually be derived by a non-ideal reasoner. In Konolige's deduction model of belief he addresses the problem of non-ideal reasoners or limited agents. His semantics was developed in an effort to define accurate models of the beliefs of robots. By considering robots the (incomplete) set of rules of the agent is available to us, the outside observers. The semantics involves a set of initial beliefs and some algorithm which can be applied to these beliefs in order to derive new beliefs. This algorithm may for example be based on a rule of inference like modus ponens which may be applied only a limited number of times. The only assumption made about the algorithm is that the agent will apply it whenever possible. By doing this the agent will arrive at all the possible inferences (from this algorithm). We therefore do not assume that the set of beliefs generated by the algorithm is closed under logical consequence but do assume that it is *deductively closed* in the sense that applying the algorithm any further will not produce new beliefs. If the algorithm is not logically complete, being deductively closed is not the same as being closed under logical consequence. Deductive closure is a much weaker condition than consequential closure. A key feature of Konolige's approach is that the deduction process of an agent can be as different as we like from those of other agents. It is even possible that the agent's set

of available rules is empty, i.e. he is equipped with the trivial algorithm with no rules of inference.

4.1 Proof Theory for Bounded Reasoning

In the definition of belief sets Konolige makes use of 'deduction structures'. These deduction structures consist of two parts namely a set of initial or base beliefs and some algorithm for generating new beliefs from old ones, typically an algorithm based on rules like modus ponens. Suppose we have some logical language L. The formal definition of a deduction structure is then as follows:

- A *deduction structure* $\mathcal{DS}$ is a pair $\langle \Gamma, \mathcal{DR} \rangle$ where Γ is a set of sentences of L, the base beliefs, and $\mathcal{DR}$ is a set of deduction rules.

What is meant by a deduction rule? Well, following Konolige, a rule of inference is *deductive* if it has the following two properties:
(i) *Provinciality*: The number of premises of the rule is fixed and finite.
(ii) *Effectiveness*: The rule is an effectively computable function of its premises.
The proof theory given in chapter 1 is an example of the use of deduction rules: modus ponens is a two-premise rule and the axioms can be thought of as zero-premise rules. An example of a rule of inference which is not provincial is 'Given any finite set Γ of sentences, infer the conjunction of the members of the set'. An example of a rule of inference which is not effective is 'Given two sentences A and B of a predicate language, infer $A \rightarrow B$ when B is a semantic consequence of A'.

- Let $\mathcal{DR}$ be the agent's set of deduction rules. Then we abbreviate the notion 'A can be derived from the set Γ by applying the rules in $\mathcal{DR}$' by $\Gamma \vdash_{\mathcal{DR}} A$.

Suppose we have an algorithm based on modus ponens but time is limited - say only n applications of the rule are permitted for some fixed n. How can such a situation be formalised? Well, one way would be to attach a label $N(k)$ to each proposition where k is the minimum number of applications of modus ponens required to infer the proposition from the set of initial beliefs. Then a deduction rule representing a suitable modification of modus ponens would be the following:
'For all k and m such that $k + m + 1 \leq n$, if the label associated with A is $N(k)$ and the label associated with $A \rightarrow B$ is $N(m)$, then B may be inferred.'
We assume closure under deduction, i.e. that the rules are applied exhaustively so that all possible conclusions are drawn. Before giving a formal definition of deductive closure we define the belief set of an agent relative to a deduction structure.

- The *belief set* of an agent relative to a deduction structure $\mathcal{DS} = \langle \Gamma, \mathcal{DR} \rangle$ is
  $bel(\mathcal{DS}) = bel(\langle \Gamma, \mathcal{DR} \rangle) = \{A \mid \Gamma \vdash_{\mathcal{DR}} A\}$.

We often abbreviate 'belief set of an agent relative to a deduction structure' to 'belief set of a deduction structure'.
Now we can define deductive closure:

- A deduction structure $\mathcal{DS} = (\Gamma, \mathcal{DR})$ is *deductively closed* iff
  $\Gamma \subseteq bel((\Gamma, \mathcal{DR}))$, and
  if $\Gamma \vdash_{\mathcal{DR}} A$ and $\Gamma, A \vdash_{\mathcal{DR}} B$, then $\Gamma \vdash_{\mathcal{DR}} B$.

Remember that we work with $\vdash_{\mathcal{DR}}$ and that the last condition does *not* state that if $\Gamma \vdash_{\mathcal{DR}} A$ and $\Gamma, A \vdash B$, then $\Gamma \vdash_{\mathcal{DR}} B$. Deductive closure guarantees that all the base beliefs of the agent are in the belief set and that any sentence derivable from the base beliefs can take part in further derivations.

Suppose $bel(\mathcal{DS})$ is the belief set of an agent. If the sentences A and $\neg A$ are both members of $bel(\mathcal{DS})$, then the belief set is inconsistent and we call it *contradictory*. (A and $\neg A$ may be base beliefs, i.e. members of Γ, or may be derived beliefs.) If, on the other hand, $A \in bel(\mathcal{DS})$ and $\neg A$ is a logical consequence of $bel(\mathcal{DS})$ but cannot be derived from the set of initial beliefs by applying the agent's deduction rules, then the set will (still) be logically inconsistent but not contradictory. This contrasts with the approaches of previous chapters where logically inconsistent belief sets were contradictory due to the logical omniscience of the agents.

We have not said anything yet about which language will be used. The deduction model of Konolige is such that nested beliefs are possible, i.e. beliefs about beliefs. Furthermore, these beliefs need not be entertained by one and the same agent, for example agent A may have beliefs about the beliefs of agent B. Konolige uses a (nonmodal) predicate language without function constants as 'internal' language and then defines a modal language, called L^B, and a variation of the modal language, called L^{Bq}. In order to focus on the essential features of the construction we will, however, first use a propositional internal language, define a modal external language and give a few examples of its use, and then consider predicate internal languages.

4.2 Semantics

Let L be some propositional language, called the internal language, and $S_1, S_2, \ldots, S_n$ be n agents. Define an external language L' based on L as follows:

- L' includes the sentences and formation rules of L and
- if A is a sentence of L, then $[S_i]A$ is a sentence of L'.

A *belief atom* is a sentence $[S_i]A$ and if a sentence of L' does not include a belief operator, it is called an *objective* or *ordinary* sentence. Suppose we want to represent beliefs about beliefs, i.e. we want to allow sentences like $[S_2][S_2]A$ and $[S_2][S_3]A$. The natural choice for an internal language would then be L' itself, i.e. 'L' in the above definition would be substituted by 'L'' and the definition would become recursive. Unless otherwise stated, however, we will take the definition as is and assume L to be a nonmodal propositional language.

Suppose we have three agents. Here are four examples of sentences of the language L' based on the propositional language L with two atoms, p and q:

$q \rightarrow p$
$\neg[S_2]q$
$p \wedge [S_3]q$
$[S_2](q \vee \neg p) \rightarrow [S_1](\neg p)$.

Informally we read a belief atom $[S_i]A$ as 'Agent S_i believes A' or, equivalently, 'A is a member of the belief set of agent S_i'. By this we mean that A is either a member of the set of base beliefs of agent S_i or is derived by the rules of the deduction structure of agent S_i. Suppose we read p as 'The green light is on' and q as 'The red light is on'. An interpretation of L' above must then be such that the sentence $[S_2](q \vee \neg p) \rightarrow [S_1](\neg p)$ will be understood as 'If agent S_2 believes that either the red light is on or the green light is not on, then agent S_1 will believe that the green light is not on'.

Consider an internal language L. Then we construct an interpretation of L' by combining an interpretation of L with a set of deduction structures, one for each agent. Suppose f is an assignment of truth values to the atoms of L. In chapter 1 an interpretation of L was defined as a valuation v_f which extended f to all the sentences of L. Let $\rho(i)$ be a function which gives the set of deduction rules of agent S_i, so $\rho(i) = \mathcal{DR}_i$. An interpretation m' of L' is called a $B(L,\rho)'$-model and it is defined as follows:

- A *$B(L,\rho)'$-model* m' of L' is a tuple $\langle f, \mathcal{DS}^* \rangle$ where f is an assignment of truth values to the atoms of the internal language L and $\mathcal{DS}^* = \{\mathcal{DS}_1, \mathcal{DS}_2, \ldots\}$ is a sequence of deduction structures $\langle \Gamma_i, \rho_i \rangle$, one for each agent, where the members of Γ_i are sentences of L.

Let $m' \models A$ abbreviate the assertion that the sentence A has truth value T under the interpretation m'. Then truth values are given to the sentences of the language L' by the following rules:

- $m' \models A$ if A is atomic and $f(A) =$T,
- $m' \models \neg A$ iff $m' \not\models A$,
- $m' \models (A \vee B)$ iff $m' \models A$ or $m' \models B$,
- $m' \models (A \wedge B)$ iff $m' \models A$ and $m' \models B$,
- $m' \models (A \rightarrow B)$ iff $m' \not\models A$ or $m' \models B$,
- $m' \models (A \leftrightarrow B)$ iff $m' \models A$ and $m' \models B$, or $m' \not\models A$ and $m' \not\models B$ and
- $m' \models [S_i]A$ iff $A \in bel(\mathcal{DS}_i)$.

Let us look at four examples.

- In the first example we take L to be the language with two atoms. Suppose we have three agents. Given $f(p) =$T and $f(q) =$F, let m' be the $B(L,\rho)'$-model consisting of f and three deduction structures. Is the sentence $p \wedge [S_3]q$ true in m'? In order to answer the question we need the deduction structure of agent S_3. Suppose

  $\mathcal{DS}_3 = \langle \Gamma_3, \mathcal{DR}_3 \rangle = \langle \{q\}, \{$ *From a sentence* $p \vee q$ *derive* $p\} \rangle$.

  We now investigate the truth of the sentence $p \wedge [S_3]q$. From the definition of f we know $m' \models p$. Is $[S_3]q$ also true in m', in other words is it the case that $q \in bel(\mathcal{DS}_3)$? Well, we see that q is a member of the set Γ_3 (the base beliefs of agent S_3), hence we conclude that $m' \models [S_3]q$. Note that this is a case of mistaken belief because q is not true in m'. The sentence $p \wedge [S_3]q$, however, is true in m'.

- For the second example we take L as the language with three atoms p, q and r. Let $f(p) = f(q) = f(r) =$T. Suppose we have only one agent. Let his base beliefs be the following set:

  $\Gamma_1 = \{p, p \rightarrow q, (p \wedge q) \rightarrow r\}$

  and let his two deduction rules be (i) modus ponens which may not be applied more than once and (ii) from A and B infer $A \wedge B$, in other words

  $\mathcal{DR}_1 =$ { *From* A *and* $A \rightarrow B$ *infer* B*, but the rule can only be applied once. From* A *and* B *infer* $A \wedge B$. }.

  Let $m' = \langle f, \mathcal{DS}_1 \rangle$ where $\mathcal{DS}_1 = \langle \Gamma_1, \mathcal{DR}_1 \rangle$. Is the sentence $[S_1]r$ true in m'? Let us see: $m' \models [S_1]r$ iff $r \in bel(\mathcal{DS}_1)$. We know r is not a base belief of agent S_1. Can it be derived? Well, we have

  $p, p \rightarrow q \vdash_{\rho_1} q$,

  so $q \in bel(\mathcal{DS}_1)$ and also

  $p, q \vdash_{\rho_1} p \wedge q$,

  so $p \wedge q \in bel(\mathcal{DS}_1)$. If the agent were able to apply modus ponens more than once, the following deduction would have been possible:

  $p \wedge q, (p \wedge q) \rightarrow r \vdash_{\rho_1} r$,

  so then r would have been a belief of agent S_1 and the sentence $[S_1]r$ would have been true in m'. This is, however, not the case and the sentence $[S_1]r$ is therefore false in m'.

- For the third example we take L as the language with two atoms, and the assignment $f(p) =$T and $f(q) =$F. Suppose we have two agents with the following deduction structures:

  $\mathcal{DS}_1 = \langle \{\}, \{$*If* A *is a belief of agent* S_2*, then* A.$\} \rangle$

  $\mathcal{DS}_2 = \langle \{q\}, \{$*From* q *infer* p.$\} \rangle$.

  Is the sentence $\neg q \wedge ([S_1]p \rightarrow q)$ true in the model constructed from the given f and the two deduction structures? Well, $\neg q$ is certainly true. For $[S_1]p \rightarrow q$ to be true $[S_1]p$ has to be false since $f(q) =$F, thus p should not be a belief of agent S_1. It is not a base belief but it can be derived:

  $q \vdash_{\rho_2} p$,

  so p is a belief of agent S_2, thus

  $[S_2]p \vdash_{\rho_1} p$.

  Hence the sentence $\neg q \wedge ([S_1]p \rightarrow q)$ will be false in m'.

- For the last example we assume we have two agents and that the language L has two atoms, p and q. Suppose we iterate the construction of L' from L and construct a new external language L'' based on L', in other words a wf like $[S_i]A$ with $A \in L'$ is a sentence of L''. An example of such a sentence is $[S_2][S_1](p \wedge q)$. Suppose $f(p) = f(q) =$T and that the two agents have the following deduction structures:

  $\mathcal{DS}_1 = \langle \{p, p \rightarrow q\}, \{$ *From* A *and* $A \rightarrow B$ *infer* B*. From* A *and* B *infer* $A \wedge B$.$\} \rangle$

  $\mathcal{DS}_2 = \langle \{\}, \{$ *If* A *is believed by agent* S_1*, infer* A*. From* A *infer* $[S_1]A$.$\} \rangle$.

Let m' be the model constructed from the given f and the two deduction structures. Is the sentence $[S_2][S_1](p \wedge q)$ true in m'? Let us see:

$m' \models [S_2][S_1](p \wedge q)$ iff $[S_1](p \wedge q) \in bel(\mathcal{D}S_2)$. We know $[S_1](p \wedge q)$ is not a base belief of agent S_2, but it can be derived:

$p, p \rightarrow q \vdash_{\rho_1} q$,

so both p and q are beliefs of agent S_1, thus

$p, q \vdash_{\rho_1} p \wedge q$.

So, because $p \wedge q$ is believed by agent S_1, $p \wedge q$ will become a belief of agent S_2 and hence

$p \wedge q \vdash_{\rho_2} [S_1](p \wedge q)$,

thus $[S_1](p \wedge q) \in bel(\mathcal{D}S_2)$, thus the sentence $[S_2][S_1](p \wedge q)$ is true in m'.

These examples illustrate the flexibility of Konolige's approach. Some very interesting results are obtained when the internal language L is not restricted to being propositional. This is taken up in the following sections.

4.3 The Language L^B

Consider a predicate language without function constants where, as in chapter 1, we follow the substitution approach to quantification. As before an interpretation w of such a language is a triple (ϕ, f, U) where U is the universe of discourse, ϕ is a mapping from the individual constants of the language to U and f is an assignment (i.e. a function from the the set of U-atoms to the set {T,F}). To refresh our memories, remember that we formed a set U' consisting of all the original individual constants together with U and that a U-atom is a ground atom containing only elements of U.

Let us introduce an example, called the 'tennis example', which will be useful for illustrating purposes:

- As internal language we use the predicate language with two predicate constants P_1^2 and P_2^1 and four individual constants c_1, c_2, c_3 and c_4. We take as our universe of discourse $U = \{Adam, Betty, Cecilia\}$. Then the set of U-atoms will be

  $\{P_1^2(Adam, Adam)$, $P_1^2(Adam, Betty)$, $P_1^2(Adam, Cecilia)$,
  $P_1^2(Betty, Betty)$, $P_1^2(Betty, Adam)$, $P_1^2(Betty, Cecilia)$,
  $P_1^2(Cecilia, Cecilia)$, $P_1^2(Cecilia, Adam)$, $P_1^2(Cecilia, Betty)$,
  $P_2^1(Adam)$, $P_2^1(Betty)$, $P_2^1(Cecilia)\}$.

  We further specify the mapping ϕ as follows: $\phi(c_1) = Adam$, $\phi(c_2) = Betty$, $\phi(c_3) = Cecilia$ and $\phi(c_4) = Betty$. Finally, let f be defined as follows: $f(P_1^2(Adam, Cecilia))$ =T, $f(P_2^1(Adam))$=$f(P_2^1(Betty))$ =T, with all other U-atoms mapped onto F. Let us agree to indicate $P_1^2(x, y)$ by $Likes(x, y)$ and $P_2^1(x)$ by $PlaysTennis(x)$. Then the assignment f maps all U-atoms onto F except $Likes(Adam, Cecilia)$, $PlaysTennis(Adam)$ and $PlaysTennis(Betty)$.

To be able to represent beliefs our language must be enriched by a modal operator or operators. We only allow the operator(s) to be placed in front of *sentences* of the relevant internal language. The enriched language is called L^B, the sentences of which are formally defined as follows:

- Given a predicate language L and one or more agents indicated by $[S_i]$, a *sentence of* L^B based on L is defined by the following rules:

  (i) L^B includes the sentences and formation rules of L and

  (ii) if A is a sentence of L, then $[S_i]A$ is a sentence of L^B.

As before a *belief atom* is a sentence $[S_i]A$ and if a sentence of L^B does not include a belief operator, it is called an *objective* or *ordinary* sentence. Suppose we want to represent beliefs about beliefs, i.e. we want to allow a sentence like $[S_2][S_3]A$. Having constructed L^B one may in that case take L^B itself as the internal language and construct the enriched language L^{BB} as external language. Unless otherwise stated, however, we will take the definition as is and assume L to be a (nonmodal) predicate language.

Suppose we have three agents. Here are four examples of sentences of the language L^B based on the predicate language given in the tennis example:

$P_2^1(c_2) \wedge [S_3]P_1^2(c_4, c_1)$	or, as agreed,	$PlaysTennis(c_2) \wedge [S_3]Likes(c_4, c_1)$
$\forall x_i P_1^2(x_i, c_3)$	or	$\forall x_i Likes(x_i, c_3)$
$\neg[S_2]\exists x_i P_2^1(x_i)$	or	$\neg[S_2]\exists x_i PlaysTennis(x_i)$
$[S_2]P_2^1(c_1) \rightarrow [S_1]P_2^1(c_3)$	or	$[S_2]PlaysTennis(c_1) \rightarrow [S_1]PlaysTennis(c_3)$.

The following formulae are not sentences of the language:

$[S_1][S_3]P_1^2(c_4, c_1)$	or	$[S_1][S_3]Likes(c_4, c_1)$
$[S_2]P_2^1(x_2)$	or	$[S_2]PlaysTennis(x_2)$
$\exists x_i[S_1]P_2^1(x_i)$	or	$\exists x_i[S_1]PlaysTennis(x_i)$.

The first formula, $[S_1][S_3]Likes(c_4, c_1)$, is not a sentence of L^B because a modal operator may only be attached to a sentence of the *internal* language. The formula $[S_2]PlaysTennis(x_2)$ is not a sentence of L^B because $PlaysTennis(x_2)$ contains a free variable and thus is not a sentence. For the same reason the third formula above namely $\exists x_i[S_1]PlaysTennis(x_i)$ does not qualify - S_1 is attached to a formula containing a free variable. This (third) formula is an example of 'quantifying-in': $\exists x_i$ in front of a modal operator followed by a formula containing the free variable x_i. This topic will be discussed in the next section.

Informally we read a belief atom $[S_i]A$ as 'Agent S_i believes A' or, equivalently 'A is a member of the belief set of agent S_i'. By this we mean that A is either a member of the set of base beliefs of agent S_i or is derived by the rules of the deduction structure of agent S_i. An interpretation of the language must be such that, in the tennis example above, the sentence $PlaysTennis(c_2) \wedge [S_3]Likes(c_4, c_1)$ will be understood as 'c_2 has property *PlaysTennis* and agent S_3 believes that c_4 and c_1 are related by property *Likes*'. It seems as if an interpretation of L^B should include an interpretation of the internal language (so that the truth or falsehood of 'c_2 has property *PlaysTennis*' can be determined) and also the base beliefs and deduction rules of all the agents (so that the truth or falsehood of 'Agent S_3 believes that c_4 and c_1 are related by property *Likes*' can be determined).

Formally we construct an interpretation of L^B by combining an interpretation of the internal language L (say) with a set of deduction structures, one for each agent. L is the internal language of all the deduction structures. Let $\rho(i)$ be a function which gives the set of deduction rules of agent S_i, so $\rho(i) = \mathcal{DR}_i$. An interpretation of L^B is called a $B(L, \rho)$-model. It is defined as follows:

- A $B(L,\rho)$-*model* m of L^B is a tuple $\langle \phi, f, U, \mathcal{DS}^*, \eta^* \rangle$ where the first three elements are an interpretation of the internal language L, $\mathcal{DS}^* = \{\mathcal{DS}_1, \mathcal{DS}_2, \ldots\}$ is a sequence of deduction structures $\langle \Gamma_i, \rho_i \rangle$, one for each agent, where the members of Γ_i are sentences of L, and $\eta^* = \{\eta_1, \eta_2, \ldots\}$ is a sequence of functions on U, one for each agent.

The sequence η^* of functions will only start playing a role when we consider quantifying-in. We ignore it in the rest of this section.

Let $m \models A$ abbreviate the assertion that the sentence A has truth value T under the interpretation m. Then truth values are given to the sentences of the language L^B by the following rules:

- $m \models A$ if A is a ground atom and $f(A^\phi) =$T where A^ϕ is the string in which every constant c in A has been replaced by $\phi(c)$, i.e. A^ϕ is the U-atom corresponding, under the mapping ϕ, to the original atom A,
- $m \models \neg A$ iff $m \not\models A$,
- $m \models (A \vee B)$ iff $m \models A$ or $m \models B$,
- $m \models (A \wedge B)$ iff $m \models A$ and $m \models B$,
- $m \models (A \rightarrow B)$ iff $m \not\models A$ or $m \models B$,
- $m \models (A \leftrightarrow B)$ iff $m \models A$ and $m \models B$, or $m \not\models A$ and $m \not\models B$,
- $m \models \forall x(A)$ iff for all $u \in U$, $m \models (A^x_u)$,
- $m \models \exists x(A)$ iff for some $u \in U$, $m \models (A^x_u)$ and
- $m \models [S_i]A$ iff $A \in bel(\mathcal{DS}_i)$.

Let us look at an example. Assume we have the language and interpretation of the tennis example and suppose the deduction structure of agent $[S_3]$ is the following:

$\mathcal{DS}_3 = \langle \Gamma_3, \mathcal{DR}_3 \rangle = \langle \{Likes(c_1, c_2), \forall x_i PlaysTennis(x_i)\}, \{$*From a sentence* $\forall x_i A$ *deduce that* $A^x_{c_2}$.$\}\rangle$.

Is the sentence $PlaysTennis(c_2) \wedge [S_3]Likes(c_4, c_1)$ true? Well, let m be the $B(L,\rho)$-model consisting of the interpretation in the tennis example and three deduction structures, only one of which, namely $\mathcal{DS}_3$, is relevant here. We now investigate the truth of the sentence $PlaysTennis(c_2) \wedge [S_3]Likes(c_4, c_1)$. First we find that $f((PlaysTennis(c_2))^\phi) = f(PlaysTennis(\phi(c_2)) = f(PlaysTennis(Betty)) =$ T. Now we must decide whether the sentence $Likes(c_4, c_1)$ is a member of the belief set of agent S_3. It is not a base belief - the only base belief of agent S_3 involving the predicate $Likes$ is $Likes(c_1, c_2)$. Is it possible to derive $Likes(c_4, c_1)$? No, because application of the only deduction rule will not produce it. Therefore the sentence $PlaysTennis(c_2) \wedge [S_3]Likes(c_4, c_1)$ is not true in the interpretation m.

The internal language, L, and the deduction rules, given by ρ^*, are fixed in a $B(L,\rho)$-model, but not the interpretation of L. We say a sentence is $B(L,\rho)$-*satisfiable* if it is true in some $B(L,\rho)$-model and $B(L,\rho)$-*valid* if it is true in all $B(L,\rho)$-models.

Let us agree that if $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ is a set of sentences, then $[S_i]\mathcal{A}$ stands for $\{[S_i]A_1, [S_i]A_2, \ldots, [S_i]A_n\}$ and $\neg[S_i]\mathcal{A}$ stands for $\{\neg[S_i]A_1, \neg[S_i]A_2, \ldots, \neg[S_i]A_n\}$. There

is an important relation between the satisfiability of sentences containing belief operators in the external language and derivations in the internal language. This is stated by the following lemma, called the *attachment lemma.*

Lemma 4.1 *The denumerable set* $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ *is* $B(L,\rho)$*-unsatisfiable iff for some* $C_j \in \mathcal{C}$ *we have* $\mathcal{A} \vdash_{\rho(i)} C_j$.

Assume first that for some sentence $C_j \in \mathcal{C}$ we have $\mathcal{A} \vdash_{\rho(i)} C_j$. Now suppose the set $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ is $B(L,\rho)$-satisfiable. This means that there must be some deduction structure $\mathcal{DS}_i = \langle \Gamma_i, \rho(i)\rangle$ such that all members of $\mathcal{A}$ are members of the belief set of agent S_i and no element of $\mathcal{C}$ is a member of the belief set of agent S_i. But, from the closure property of deduction structures, we know that C_j must be in the agent's belief set. From this contradiction we thus conclude that $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ is $B(L,\rho)$-unsatisfiable.

Conversely, assume $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ is $B(L,\rho)$-unsatisfiable. Now suppose for all sentences $C_j \in \mathcal{C}$ we have that $\mathcal{A} \nvdash_{\rho(i)} C_j$. Then we can construct a deduction structure $\mathcal{DS}_i = \langle \mathcal{A}, \rho(i)\rangle$ which is such that no member of $\mathcal{C}$ is in $bel(\mathcal{DS}_i)$. So for the $B(L,\rho)$-model $m = \langle \phi, f, U, \{\ldots, \mathcal{DS}_i, \ldots\}, \eta^*\rangle$ we will have that $m \Vdash [S_i]A_j$ for every $A_j \in \mathcal{A}$ and $m \nVdash [S_i]C_j$ for every $C_j \in \mathcal{C}$, thus $m \Vdash \neg[S_i]C_j$ for every $C_j \in \mathcal{C}$. But this contradicts the unsatisfiability assumption, hence for some sentence $C_k \in \mathcal{C}$ we must have $\mathcal{A} \vdash_{\rho(i)} C_k$. ♠

The lemma is called the attachment lemma from the way it attaches the question of satisfiability in the external language L^B to the derivation process in the internal language L.

It is possible to define *classes of* $B(L,\rho)$*-models.* Some of the classes are defined by the form of the deduction rules and some by placing conditions on the belief sets. The latter will be indicated by specifying certain schemas that are characteristic of the class: all instances of the schema will be satisfied by every model in the class and every model not in the class will falsify some instance of the schema. We close this section by discussing a few classes.

Saturation. We call a deduction structure *saturated* iff the set of deduction rules is sound and complete with respect to the semantics of the internal language. A class of $B(L,\rho)$-models is called saturated if ρ is such that it associates with every agent a sound and complete set of deduction rules. The belief sets of saturated deduction structures are closed under logical consequence, in other words in these cases we have ideal reasoners. This is stated by the following theorem.

Theorem 4.1 *Let* ρ *be such that the class of* $B(L,\rho)$*-models is saturated. Let* m *be a member of the class and* $\mathcal{A}$ *a set of sentences from* L. *If the sentence* C *of* L *is a logical consequence of* $\mathcal{A}$ *(i.e.* $\mathcal{A} \models C$*), then for all agents* S_i *we have that* $m \Vdash [S_i]\mathcal{A} \rightarrow [S_i]C$.

Suppose $\mathcal{A} \models C$. Then $\mathcal{A} \vdash_{\rho(i)} C$ because the deduction structures are saturated. By lemma 4.1 it follows that if all members of $\mathcal{A}$ are in the belief set of agent S_i, then C will also be a member of the belief set. ♠

Corollary 4.1 *Given a saturated class of* $B(L,\rho)$*-models, we have that*
(i) for every valid sentence A *of the language* L, $[S_i]A$ *is* $B(L,\rho)$*-valid and*
(ii) the sentence $[S_i](A \rightarrow C) \rightarrow ([S_i]A \rightarrow [S_i]C)$ *is* $B(L,\rho)$*-valid.*
The results follow immediately from the above theorem. ♠

Thus the belief sets of ideal reasoners will contain all valid sentences of the language

and will be closed under modus ponens. In the case of a non-ideal reasoner, valid sentences may be absent from his belief set and even if both A and $A \rightarrow B$ are members of his belief set, B may be absent (for example if the set of deduction rules is empty).

Knowledge. Let us consider knowledge as true belief, i.e. let us consider the class consisting of just those $B(L,\rho)$-models in which the agent's beliefs agree with the facts of the 'outside' world. A $B(L,\rho)$-model $m = \langle \phi, f, U, \mathcal{DS}^*, \eta^* \rangle$ is in this class iff for every deduction structure $\mathcal{DS}_i \in \mathcal{DS}^*$, if $A \in bel(\mathcal{DS}_i)$, then $m \models A$. The schema which characterises this class is

$[S_i]A \rightarrow A$.

Noncontradiction. We have seen that a belief set may be inconsistent and contradictory (containing some sentence A and also $\neg A$) or inconsistent without being contradictory. Consider the class consisting of just those $B(L,\rho)$-models in which no belief set is contradictory. A $B(L,\rho)$-model $m = \langle \phi, f, U, \mathcal{DS}^*, \eta^* \rangle$ belongs to this class iff for every deduction structure $\mathcal{DS}_i \in \mathcal{DS}^*$, if $A \in bel(\mathcal{DS}_i)$, then $\neg A \notin bel(\mathcal{DS}_i)$. The schema which characterises this class is

$[S_i]A \rightarrow \neg[S_i]\neg A$.

Note that the belief set of an agent whose deduction structure is both noncontradictory and saturated will be consistent.

Introspection. Suppose the internal language L contains a belief operator $[S_i]$. A $B(L,\rho)$-model $m = \langle \phi, f, U, \mathcal{DS}^*, \eta^* \rangle$ is in the positive introspection class iff for every deduction structure $\mathcal{DS}_i \in \mathcal{DS}^*$, if $A \in bel(\mathcal{DS}_i)$, then $[S_i]A \in bel(\mathcal{DS}_i)$. The schema which characterises this class is

$[S_i]A \rightarrow [S_i][S_i]A$.

Similarly, a $B(L,\rho)$-model m is in the negative introspection class iff for every deduction structure $\mathcal{DS}_i \in \mathcal{DS}^*$, if $A \notin bel(\mathcal{DS}_i)$, then $\neg[S_i]A \in bel(\mathcal{DS}_i)$. The schema which characterises this class is

$\neg[S_i]A \rightarrow [S_i]\neg[S_i]A$.

4.4 Quantifying-in and Naming Maps

We have seen that $\exists x_i[S_1]PlaysTennis(x_i)$ was an example of an impermissible wf of the language L^B, the reason being that a modal operator may only be attached to a sentence of the internal language. So the notion that agent S_1 believes a certain individual to have the property of $PlaysTennis$ cannot be expressed in the language L^B. Note that the above is not the same as saying that agent S_1 believes merely in the existence of somebody who has the property $PlaysTennis$. The latter notion can be expressed, namely by the sentence $[S_1]\exists x_i PlaysTennis(x_i)$. In (nonmodal) predicate logic a sentence $\exists x A$ is satisfiable iff the sentence A_c^x is satisfiable where c is a Skolem constant not occurring in A. If we, however, try to apply the same technique here we will get in both cases the sentence $[S_1]PlaysTennis(c)$ which is not satisfactory - sentences with very different meanings are transformed into the same sentence.

What is actually meant by the two sentences above? Well, the second sentence, namely $[S_1]\exists x_i PlaysTennis(x_i)$, is understood to mean that $\exists x_i PlaysTennis(x_i)$ is a member of the belief set of agent S_1. The first sentence, namely $\exists x_i[S_1]PlaysTennis(x_i)$, is under-

stood to mean that there exists an individual such that agent S_1 believes *this particular* individual to play tennis. So in this case agent S_1 has the sentence $PlaysTennis(c)$ in his belief set where c represents agent S_1's way of identifying the relevant individual, i.e. c is an *id constant* whose intended denotation is the appropriate value of x_i. (Konolige uses the term 'id constant' as an abbreviation for 'identifiable constant'. These constants are, loosely speaking, computationally significant because they contain all the information necessary to identify the relevant individual.) Thus the sentences $[S_1]PlaysTennis(c)$ and $\exists x_i[S_1]PlaysTennis(x_i)$ say the same as long as c refers to the appropriate value of x_i. This is where the functions η_i in a model $m = \langle \phi, f, U, \mathcal{DS}^*, \eta^* \rangle$ come in. Every agent S_i associates every member of the universe of discourse U with some constant.

- The function η_i maps U onto a subset of the individual constants of the language. The members of the subset are called the *id constants* of agent S_i and we call the function η_i a *naming map.*

We may allow η_i to be a partial map when agent S_i does not know every individual in U. We may even let η_i be a relation rather than a function if agent S_i associates more than one id constant with the same individual. Unless otherwise stated, however, we take η_i as a total function.

Let m be the (L, ρ)-model consisting of the interpretation in the tennis example and a deduction structure for agent S_1. Intuitively $m \models \exists x_i[S_1]PlaysTennis(x_i)$ should be true iff $[S_1]PlaysTennis(u)$ is true for some $u \in U$. Further, $[S_1]PlaysTennis(u)$ is true iff there is a sentence in the belief set of agent S_1 asserting that u has property $PlaysTennis$. Unfortunately sentences of the belief set are sentences of the language L, thus $PlaysTennis(u)$ cannot be in any belief set. However, if L has an id constant denoting u, a sentence that expresses the same information could well be in the belief set, namely the sentence $PlaysTennis(c)$ where $\eta_1(u) = c$.

Suppose agent S_1 has the following naming map: $\eta_1(Adam) = c_1$, $\eta_1(Betty) = c_4$, $\eta_1(Cecilia) = c_3$. Suppose further that his deduction structure is the following:

$\mathcal{DS}_1 = \langle \{Likes(c_1, c_4), PlaysTennis(c_3)\}, \{\} \rangle$.

Then the sentence $\exists x_i[S_1]PlaysTennis(x_i)$ will be true in the model m: take $u \in U$ as $Cecilia$, then $\eta_1(Cecilia) = c_3$ and we have $PlaysTennis(c_3)$ as a (base) belief of agent S_1. (Note that this is a misplaced belief because $PlaysTennis(c_3)$ is not true in m.)

The naming map must be defined in such a way that the function ϕ always takes the id constant of an individual back to that specific individual. By this we mean that for all agents S_i and all $u \in U$, we have that $\phi(\eta_i(u)) = u$. So the naming map is the partial inverse of the function ϕ. It is, however, not necessarily the case that $\eta_i(\phi(c)) = c$. In the example above we have

$\phi(\eta_1(Adam)) = \phi(c_1) = Adam$,
$\phi(\eta_1(Betty)) = \phi(c_4) = Betty$, and
$\phi(\eta_1(Cecilia)) = \phi(c_3) = Cecilia$,

but

$\eta_1(\phi(c_2)) = \eta_1(Betty) = c_4 \neq c_2$.

In the special case where η is the inverse of ϕ the id constant is called a *standard name.*

4.5 The language L^{Bq}

We now extend the language L^B by permitting quantifying in and introducing an additional operator. Suppose the internal language L contains constants. Then the language $L^\bullet$ is

identical to L except that the number of constants is doubled by having $\bullet c_i$ as an additional constant for every individual constant c_i. In order to get an external language we extend the language L^B to the language L^{Bq}, formally defined as follows:

- Given an internal language L and one or more agents indicated by $[S_i]$, a *wf of* L^{Bq} based on L is defined recursively by the following rules:

  (i) L^{Bq} includes the sentences and formation rules of L and

  (ii) if A is a wf of $L^\bullet$, then $[S_i]A$ is a wf of L^{Bq}.

We see that in the language L^{Bq} wfs of the language $L^\bullet$ are used as the arguments of belief atoms while in L^B we had sentences of L as the arguments.

Let us look at an example. Let L be the language given in the tennis example. The language $L^\bullet$ is identical to L except that the individual constants are now $c_1, \bullet c_1, c_2, \bullet c_2, c_3, \bullet c_3, c_4, \bullet c_4$. Suppose we have three agents. Here are a few sentences of the language L^{Bq} based on L:

$PlaysTennis(c_2) \wedge [S_3]Likes(c_4, \bullet c_1)$

$\forall x_i[S_1]Likes(x_i, c_3)$

$\neg[S_2]\exists x_i PlaysTennis(x_i) \vee \exists x_i[S_1]Likes(\bullet c_2, x_i)$

$[S_2]PlaysTennis(\bullet c_1) \rightarrow [S_1]PlaysTennis(c_3)$.

The following formulae are not sentences of the language L^{Bq}:

$[S_1][S_3]Likes(c_4, \bullet c_1)$

$PlaysTennis(\bullet c_2)$,

the first formula because $[S_1]$ may only be attached to a formula of the language $L^\bullet$ and the second because bullet operators may only appear in belief atoms.

We define an interpretation of the language L^{Bq} in exactly the same way as an interpretation of L^B namely as a tuple $m = \langle \phi, f, U, \mathcal{DS}^*, \eta^* \rangle$. The difference comes in when we give the rules by which truth values are given to the sentences of the language. As in the case of the language L^B we want to follow a substitution approach to quantification. In other words, we want the truth of $\exists x[S_i]A$ to be determined by the truth of $([S_i]A)^x_u$ for some $u \in U$. However, the truth of $([S_i]A)^x_u$ should in turn be determined by whether $A^x_{\eta_i(u)} \in bel(\mathcal{DS}_i)$. As substitution of terms for variables is of general importance (for example, if one adopts a proof theory involving quantifier elimination), and as the substitution of Skolem constants is inadequate to cope with quantifying-in, we define substitution in such a way that the id constant $\bullet c$ is substituted instead of c when necessary. Formally we proceed as follows:

Let A be any wf of L^{Bq}. For every variable x and individual constant c the formula A^x_c is given by the following rules:

- If A is an atom, then A^x_c is the result of substituting c for every occurrence of x in A.
- For any wf A, $(\neg A)^x_c = \neg A^x_c$.
- For any wfs A and B, $(A \odot B)^x_c = (A^x_c \odot B^x_c)$ where $\odot \in \{\vee, \wedge, \rightarrow, \leftrightarrow\}$.
- For any wf A, $(Qx(A))^x_c = Qx(A)$ where $Q \in \{\forall, \exists\}$.
- For any wf A, $(Qy(A))^x_c = Qy(A^x_c)$ where $Q \in \{\forall, \exists\}$ and $x \neq y$.

- For any wf A,

$$\begin{aligned}([S_i]A)^x_c &= [S_i]A^x_{\bullet c} \quad \text{for } c \text{ a constant of } L\\ ([S_i]A)^x_c &= [S_i]A^x_c \quad \text{for } c \in U.\end{aligned}$$

For a set Γ of wfs we sometimes use the notation Γ^x_c by which is meant that x is replaced by c in all members of Γ according to the rules above.

Given the above rules, what will be a sensible way of allocating truth values to sentences of the language L^{Bq}? As before we generalize the definition of substitution and take A^ϕ to be the string that results when all constants c in A have been replaced by $\phi(c)$. Analogously we define A^{η_i} as A with all occurrences of $u \in U$ replaced by $\eta_i(u)$ and all occurrences of $\bullet c$ replaced by $\eta_i(\phi(c))$. (Note that if c is a standard name, $\eta_i(\phi(c)) = c$.) Then the rules for giving truth values are the following:

- $m \mid\!\vdash A$ iff A is a ground atom and $f(A^\phi) = \mathrm{T}$ where A^ϕ is the U-atom corresponding, under the mapping ϕ, to the original atom A as explained above,
- $m \mid\!\vdash \neg A$ iff $m \not\mid\!\vdash A$,
- $m \mid\!\vdash (A \vee B)$ iff $m \mid\!\vdash A$ or $m \mid\!\vdash B$,
- $m \mid\!\vdash (A \wedge B)$ iff $m \mid\!\vdash A$ and $m \mid\!\vdash B$,
- $m \mid\!\vdash (A \rightarrow B)$ iff $m \not\mid\!\vdash A$ or $m \mid\!\vdash B$,
- $m \mid\!\vdash (A \leftrightarrow B)$ iff $m \mid\!\vdash A$ and $m \mid\!\vdash B$, or $m \not\mid\!\vdash A$ and $m \not\mid\!\vdash B$,
- $m \mid\!\vdash \forall x(A)$ iff for all $u \in U$, $m \mid\!\vdash A^x_u$,
- $m \mid\!\vdash \exists x(A)$ iff for some $u \in U$, $m \mid\!\vdash A^x_u$ and
- $m \mid\!\vdash [S_i]A$ iff $A^{\eta_i} \in bel(\mathcal{DS}_i)$.

Let us look at an example, using the internal language L and the interpretation (ϕ, f, U) of the tennis example. Suppose we have three agents and are interested in agent S_3. We assume the deduction structure of agent S_3 is

$\mathcal{DS}_3 = \langle \{Likes(c_1, c_3), PlaysTennis(c_3)\}, \{\}\rangle$

and the naming map is $\eta_3(Adam) = c_1$, $\eta_3(Betty) = c_4$, $\eta_3(Cecilia) = c_3$. Is the sentence $PlaysTennis(c_2) \wedge [S_3]Likes(c_4, \bullet c_1)$ true in $m = \langle \phi, f, U, \mathcal{DS}^*, \eta^*\rangle$?

The truth value of the first part of the sentence is easily determined: $(PlaysTennis(c_2))^\phi = PlaysTennis(\phi(c_2)) = PlaysTennis(Betty)$ which is mapped onto T by f. In order to find the truth value of $[S_3]Likes(c_4, \bullet c_1)$ we have to investigate whether $(Likes(c_4, \bullet c_1))^{\eta_3}$ is a member of $bel(\mathcal{DS}_3)$. According to the definition of A^{η_i} we have that $(Likes(c_4, \bullet c_1))^{\eta_3} = Likes(c_4, \eta_3(\phi(c_1))) = Likes(c_4, \eta_3(Adam)) = Likes(c_4, c_1)$ which is not a member of the belief set of agent S_3 and thus the sentence is not true in m.

We look at a second example, this time involving quantifying-in. Is $\exists x_i[S_3]Likes(\bullet c_1, x_i)$ true in m? Following the rules we have to find out

whether there exists a $u \in U$ such that $m \mid\!\vdash ([S_3]Likes(\bullet c_1, x_i))^{x_i}_u$,
i.e. whether there exists a $u \in U$ such that $m \mid\!\vdash [S_3](Likes(\bullet c_1, x_i))^{x_i}_u$,
i.e. whether there exists a $u \in U$ such that $m \mid\!\vdash [S_3]Likes(\bullet c_1, u)$,
i.e. whether there exists a $u \in U$ such that $(Likes(\bullet c_1, u))^{\eta_3} \in bel(\mathcal{DS}_3)$,
i.e. whether there exists a $u \in U$ such that $Likes(\eta_3(\phi(c_1)), \eta_3(u)) \in bel(\mathcal{DS}_3)$,

i.e. whether there exists a $u \in U$ such that $Likes(\eta_3(Adam), \eta_3(u)) \in bel(\mathcal{DS}_3)$,

i.e. whether there exists a $u \in U$ such that $Likes(c_1, \eta_3(u)) \in bel(\mathcal{DS}_3)$.

We know that $\eta_3(Cecilia) = c_3$ and that $Likes(c_1, c_3)$ is a (base) belief of agent S_3. Thus there does exist a $u \in U$ (namely *Cecilia*) such that $Likes(c_1, \eta_3(u))$ is in the belief set of agent S_3 and we (at last) are able to conclude that the sentence $\exists x_i[S_3]Likes(\bullet c_1, x_i)$ is true in m.

The substitution and satisfaction rules above are defined in such a way that satisfiability is preserved in the sense stated in the following theorem.

Theorem 4.2 *Given a language L^{Bq}, an interpretation m with universe U and a wf A with n free variables $x_1, x_2, \ldots, x_n$. Suppose $u_i \in U, 1 \leq i \leq n$, and individual constants $c_i, 1 \leq i \leq n$, are such that $\phi(c_i) = u_i$. Then $m \models A^{x_i}_{u_i}$ iff $m \models A^{x_i}_{c_i}$*

It is sufficient to show that the theorem holds for $n = 1$ and we use (unsubscripted) x, c and u. The theorem holds for the case where A is a ground atom because in that case we have $m \models A^x_c$ iff $f((A^x_c)^\phi) = \mathrm{T}$ iff $f(A^x_{\phi(c)}) = \mathrm{T}$ iff $f(A^x_u) = \mathrm{T}$ iff $m \models A^x_u$. For ordinary wfs the proof is by induction on the subformulae of A. Suppose now A is of the form $[S_i]B$ where B is an ordinary wf. We know $([S_i]B)^x_c = [S_i](B^x_{\bullet c})$, so we have to show that $(B^x_u)^{\eta_i} = (B^x_{\bullet c})^{\eta_i}$. But this is the case because we replace all occurrences of u on the lefthand side by $\eta_i(u)$ and all occurrences of $\bullet c$ on the righthand side by $\eta_i(\phi(c)) = \eta_i(u)$. Using induction we conclude that the theorem holds for every wf of the language. ♠

We have seen that the attachment lemma (lemma 4.1) associates satisfiability in L^B with the derivation process in L. Is it possible to formulate a similar attachment lemma for the language L^{Bq}? The answer is yes but some preliminary definitions are necessary. The problem is the presence of the bullet operator. A sentence of L^{Bq} such as $P^1_2(\bullet c)$ is, for example, not a wf of the internal language. Such a sentence has to be replaced by some suitable wf from L. But what wf will be 'suitable'? Well, we know that $(P^1_2(\bullet c))^{\eta_i} = P^1_2(\eta_i(\phi(c)))$, so $\bullet c$ refers to the id constant assigned by the naming map η_i to the individual denoted by c. Because we do not actually *know* which individual or id constant this is, it seems reasonable to uniformly substitute any arbitrary id constant for $\bullet c$ when we apply the attachment rule. There are, however, two reservations.

It is impermissible to use an id constant that is already present in the L^{Bq} sentence. For example, suppose we have the sentence

$[S_i]P^1_2(\bullet c_1) \rightarrow [S_i]P^1_2(c_2)$

where c_2 is an id constant. Then we cannot replace $\bullet c_1$ by c_2 since it is certainly the case that

$\{P^1_2(c_2)\} \vdash_{\rho(i)} P^1_2(c_2)$

but the original sentence may not be valid.

The second reservation depends on the way the agent's deduction rules handle id constants. Suppose, for example, that agent S_1 has a rule which he only applies to *Betty* namely

If $PlaysTennis(c_2)$ *then* $\forall x_i Likes(x_i, c_2)$.

where c_2 is an id constant for *Betty*. Now consider the L^{Bq} sentence

$[S_1]PlaysTennis(\bullet c) \rightarrow [S_1]\forall x_i Likes(x_i, \bullet c)$.

If an arbitrary new id constant may be substituted for $\bullet c$ and if c_2 is chosen, it will be the case that

$PlaysTennis(c_2) \vdash_{\rho(1)} \forall x_i Likes(x_i, c_2)$

and we will expect the original sentence to be valid which it is not (just pick the denotation of c to be, for example, *Adam*).

We therefore need id constants that act like any other id constant with respect to the agent's deduction rules. These id constants are called schematic constants and the formal definition is the following:

- Given a language L, let $\rho(i)$ be the deduction rules of agent S_i and c an id constant. Further, let Γ be a set of sentences and B be a sentence of L possibly containing c, such that $\Gamma \vdash_{\rho(i)} B$. Then c is a *schematic constant of* $\rho(i)$ if, for every other id constant c', we have that $\Gamma^c_{c'} \vdash_{\rho(i)} B^c_{c'}$.

The deduction rules of the agent determine which id constants are schematic constants. Any id constant that is treated in a special way (like c_2 in the example above where certain rules apply only to it) cannot be a schematic constant. Schematic constants can be applied freely in any derivation. Normally we require that the internal language contains countably many such constants.

Before we can give the attachment lemma for L^{Bq} we have to define the bullet deletion transform of a set of sentences from $L^\bullet$:

- Given a language L and a set Γ of sentences of the language $L^\bullet$, the *bullet deletion transform* of Γ namely $\Gamma^\bullet$ is the set formed by uniformly replacing every bullet constant $\bullet a_k$ of Γ with a schematic constant c_k not already in Γ where $c_i \neq c_j$ for $i \neq j$.

The bullet deletion transform of a single sentence A is written as $A^\bullet$. Now the *attachment lemma* for the language L^{Bq} can be stated.

Lemma 4.2 *The denumerable set* $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ *is* $B(L,\rho)$*-unsatisfiable iff for some* $C_j \in \mathcal{C}$ *we have* $\mathcal{A}^\bullet \vdash_{\rho(i)} C_j^\bullet$ *for all bullet deletion transforms of* $\mathcal{A}$ *and* $\mathcal{C}$.

'If' direction: We will assume that the naming map η_i of agent S_i is total. Now, assume that for some sentence $C_j \in \mathcal{C}$ we have $\mathcal{A}^\bullet \vdash_{\rho(i)} C_j^\bullet$. Suppose the set $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ is satisfiable. Then there is a model m which satisfies it. Let C_j' be the sentence of L that is denoted by C_j in m, i.e. all bullet constants $\bullet c_k$ in C_j are replaced by $\eta_i(\phi(c_k))$, and let $\mathcal{A}'$ be the set of sentences denoted by $\mathcal{A}$ in m, i.e. all bullet constants $\bullet c_k$ in the sentences of $\mathcal{A}$ are replaced by $\eta_i(\phi(c_k))$. Further, let $\mathcal{DS}_i$ be the deduction structure of agent S_i in m. Then it is the case that $C_j' \notin bel(\mathcal{DS}_i)$ and $\mathcal{A}' \subseteq bel(\mathcal{DS}_i)$. But since $\mathcal{A}^\bullet \vdash_{\rho(i)} C_j^\bullet$, it is also the case that $\mathcal{A}' \vdash_{\rho(i)} C_j'$ by the definition of schematic constants. Thus because the deduction rules are deductively closed $C_j' \in bel(\mathcal{DS}_i)$. From this contradiction we conclude that the set $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ is unsatisfiable.

'Only if' direction: To prove the lemma in the opposite direction we assume that the set $\{[S_i]\mathcal{A}, \neg[S_i]\mathcal{C}\}$ is unsatisfiable. Given a bullet deletion transform of $\mathcal{A}$ and $\mathcal{C}$, assume that for all sentences $C \in \mathcal{C}$ we have $\mathcal{A}^\bullet \nvdash_{\rho(i)} C^\bullet$. We will now construct a model m that satisfies the set, thus arriving at a contradiction. Choose an arbitrary f and a denumerably infinite U and define ϕ in such a way that each constant is mapped onto a different individual of U. Suppose the bullet deletion transform converts $\bullet a_j$ to the schematic constant c_j for agent S_i. Define the agent's naming map η_i in such a way that $\eta_i(\phi(a_j)) = c_j$. Now it will be the case that $m \models \mathcal{A}$ iff $m \models \mathcal{A}^\bullet$ and $m \models \mathcal{C}$ iff $m \models \mathcal{C}^\bullet$. Then we construct the deduction structure $\mathcal{DS}_i = \langle \mathcal{A}^\bullet, \rho(i) \rangle$ which will have the property that no member of $\mathcal{C}^\bullet$ will be in the belief set of agent S_i because of the assumption that it cannot be derived. Then, for the model $m = \langle \phi, f, U, \{\ldots, \mathcal{DS}_i, \ldots\}, \eta^* \rangle$, we have $m \models [S_i]A$ for every $A \in \mathcal{A}$ and $m \not\models [S_i]C$

for every $C \in \mathcal{C}$. This is the contradiction since we assumed that no such model exists. ♠

The assumption that the naming map η_i is a total function was needed so that C'_j would be defined in every model. If η_i is a partial function, C'_j may be undefined in some model and the set $\{[S_i]\mathcal{A}, \neg[S_i]C_j\}$ would be satisfiable even though the derivation $\mathcal{A}^\bullet \vdash_{\rho(i)} C_j^\bullet$ holds. Fortunately it is possible to preserve the attachment lemma by adding the following condition to the *'If' direction*:

All bullet constants of C_j are also in $\mathcal{A}$.

4.6 Some Properties of Quantifying-in

Suppose agent S_j has a rule *From* $P_1^1(c)$ *infer* $\exists x_i P_1^1(x_i)$ *where* c *is a schematic constant*, in other words the agent is capable of 'existential generalisation'. Then the following holds for any $B(L,\rho)-$model m:

$m \Vdash \exists x_i[S_j]P_1^1(x_i) \rightarrow [S_j]\exists x_i P_1^1(x_i)$.

Why? By the rules above $\exists x_i[S_j]P_1^1(x_i)$ is true only if there exists a $u \in U$ such that $m \Vdash [S_j]P_1^1(u)$, i.e. if there exists a $u \in U$ such that $(P_1^1(u))^{\eta_j} \in bel(\mathcal{DS}_j)$, i.e. if there exists a $u \in U$ such that $P_1^1(\eta_j(u)) \in bel(\mathcal{DS}_j)$, i.e. if $P_1^1(c') \in bel(\mathcal{DS}_j)$ where c' is some id constant. So, suppose for some id constant c', $P_1^1(c')$ is in the belief set of agent S_j. Then, because the agent is capable of existential generalisation, this means that $\exists x_i P_1^1(x_i)$ will also be in the belief set.

The above sentence will also be valid in all saturated models because existential generalisation is a sound rule. Thus, if an ideal reasoner believes that a particular individual has a certain property, he will also believe that someone has the property.

What about the converse, i.e. is it the case that the following holds for any $B(L,\rho)-$model m:

$m \Vdash [S_j]\exists x_i P_1^1(x_i) \rightarrow \exists x_i[S_j]P_1^1(x_i)$?

The answer is no. Why? Well, if $m \Vdash [S_j]\exists x_i P_1^1(x_i)$ it means $(\exists x_i P_1^1(x_i))^{\eta_j} = \exists x_i P_1^1(x_i) \in bel(\mathcal{DS}_j)$. Suppose this is the case, in other words the agent believes someone has property P_1. Then it is not necessarily the case that the consequent is true in m, because for $\exists x_i[S_j]P_1^1(x_i)$ to be satisfied by m we must have that for some $u \in U$, $[S_j]P_1^1(u)$ is true in m, i.e. for some $u \in U$, $(P_1^1(u))^{\eta_j} \in bel(\mathcal{DS}_j)$, i.e. for some $u \in U$, $P_1^1(\eta_j(u)) \in bel(\mathcal{DS}_j)$, i.e. for some id constant c, $P_1^1(c) \in bel(\mathcal{DS}_j)$ and this is not (necessarily) the case. So, if the agent believes that someone has a certain property, he will not necessarily believe that a particular individual has that property. This seems quite reasonable.

Let us consider the sentence $\forall x_i[S_j]P_1^1(x_i) \rightarrow [S_j]\forall x_i P_1^1(x_i)$. Is this sentence or its converse valid in all saturated $B(L,\rho)$-models? The answer is that it may be the case but under very specific circumstances. Note that an ideal reasoner will be able to derive from $\forall x_i P_1^1(x_i)$ that $P_1^1(c)$ for all c in L, in other words if $\forall x_i P_1^1(x_i)$ is in the belief set of agent S_j, $P_1^1(c)$ will also be in his belief set for all c. Now let us first consider the converse of the above sentence, namely

$[S_j]\forall x_i P_1^1(x_i) \rightarrow \forall x_i[S_j]P_1^1(x_i)$.

Suppose $\forall x_i P_1^1(x_i) \in bel(\mathcal{DS}_j)$. When will the consequent be true? Well, it will be true if for all $u \in U$ we have that $m \Vdash [S_j]P_1^1(u)$, i.e. if for all $u \in U$ we have $(P_1^1(u))^{\eta_j} \in bel(\mathcal{DS}_j)$, i.e. if for all $u \in U$ we have $P_1^1(\eta_j(u)) \in bel(\mathcal{DS}_j)$. So, *if* every $u_i \in U$ has an associated id constant c_i, i.e. *if* η_i is total, then the sentence will be valid.

Now consider the sentence

$\forall x_i[S_j]P_1^1(x_i) \rightarrow [S_j]\forall x_i P_1^1(x_i)$.
The antecedent is true in a model m only if, in the first place, the naming map is total (so we assume η_i to be total) and if, in the second place, for all $u_i \in U$ we have $(P_1^1(u_i))^{\eta_j} \in bel(\mathcal{DS}_j)$, i.e. if for all $u_i \in U$ we have $P_1^1(\eta_j(u_i)) \in bel(\mathcal{DS}_j)$. Suppose $\eta_j(u_i) = c_i$ where the u_i are distinct but the c_i need not be. Now assume each of these $P_1^1(c_i)$ is in the belief set of agent S_j. The consequent will be true in m if $(\forall x_i P_1^1(x_i))^{\eta_j} \in bel(\mathcal{DS}_j$. This will be the case if the agent is able to derive $\forall x_i P_1^1(x_i)$ and this will be possible only if the c_i above cover all constants of L.

4.7 Proof Methods

As proof method Konolige uses analytic tableaux with so-called signed formulae. A *signed formula* is a wf preceded by either 'T' or 'F'. In order to show that a sentence B follows from a set $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ of sentences we start a tableau with FB and TA_1, TA_2, ...,TA_n. (These formulae form the 'root' of the tableau.) By following certain rules other signed formulae are then produced. It is possible that the tableau will grow into sub-tableaux or branches. If we succeed in arriving at a contradiction, i.e. having for some sentence C both TC and FC in a branch, we *close* the branch. If all the branches are closed, we conclude that it is indeed the case that B follows from $\mathcal{A}$. It will namely be impossible to construct an interpretation where B is false and all the members of $\mathcal{A}$ are true, thus the root cannot be satisfied.

Under what conditions is the analytic tableau method deductively closed? Recall that a deduction stucture is deductively closed iff, in the first place, $\Gamma \subseteq bel(\langle \Gamma, \mathcal{DR} \rangle)$, and, in the second place, if $\Gamma \vdash_{\mathcal{DR}} A$ and $\Gamma, A \vdash_{\mathcal{DR}} B$, then $\Gamma \vdash_{\mathcal{DR}} B$. The first condition is obviously satisfied by analytic tableaux and for the second we specify the following:

- If the two tableaux respectively headed by TΓ, FA and TΓ, TA, FB both close, then the tableau headed by TΓ, FB also closes.

The rules listed below are such that this condition holds.

Before giving the actual rules let us illustrate the process with an example. We show that $A \vee (\neg A \wedge B) \vee \neg B$ is a tautology where A and B are sentences of a nonmodal language. We construct the following tableau:

(1)	F$A \vee (\neg A \wedge B) \vee \neg B$	
(2)	FA	(1)
(3)	F$\neg A \wedge B$	(1)
(4)	F$\neg B$	(1)
(5)	TB	(4)

(6)	F$\neg A$	(3)	(8)	FB	(3)
(7)	TA	(6)		$\times$	(5,8)
	$\times$	(2,7)			

The numbers on the left are the line numbers of the tableau and those on the right indicate the (previous) lines from which the current one was derived. Loosely speaking, lines (2), (3) and (4) follow from line (1) since all disjuncts of a false disjunction are false and line (5) is derived from line (4) by observing that, if $\neg B$ is false, then B must be true. Line (3) causes a split in the tableau (lines (6) and (8)) because either conjunct of a false conjunction may

be false, i.e. we have two possibilities. When a split occurs two sub-tableaux appear next to each other. Line (7) follows from line (6) where we (again) had a false negation. The symbol '×' indicates that a branch is closed. We see that both branches of the tableau are closed, thus the given sentence is a tautology.

Formally we proceed as follows. Assume we have a predicate language L, the internal language of agents $S_i, i = 1, 2, \ldots, n$. Let us now list the rules for deriving signed formulae from given signed formulae. They are grouped under four headings. Let A and B be any sentence of L.

- By a *conjunctive type* rule, if α appears in the tableau, both α_1 and α_2 are added.

α	α_1	α_2
$\mathrm{T}A \wedge B$	$\mathrm{T}A$	$\mathrm{T}B$
$\mathrm{F}A \vee B$	$\mathrm{F}A$	$\mathrm{F}B$
$\mathrm{F}A \rightarrow B$	$\mathrm{T}A$	$\mathrm{F}B$
$\mathrm{T}\neg A$	$\mathrm{F}A$	
$\mathrm{F}\neg A$	$\mathrm{T}A$	

The second rule of this type was applied in lines (2), (3) and (4) of the example above and the fifth rule was applied in lines (5) and (7).

- By a *disjunctive type* rule, if α appears in the tableau, the tableau splits into two branches, one headed by α_1 and the other by α_2.

α	α_1	α_2
$\mathrm{F}A \wedge B$	$\mathrm{F}A$	$\mathrm{F}B$
$\mathrm{T}A \vee B$	$\mathrm{T}A$	$\mathrm{T}B$
$\mathrm{T}A \rightarrow B$	$\mathrm{F}A$	$\mathrm{T}B$

The first rule of this type was applied in lines (6) and (8) above.

- By a *universal type* rule a wf A may be instantiated with an arbitrary constant.

α	$\alpha(c)$
$\mathrm{T}\forall x A$	$\mathrm{T}A_c^x$
$\mathrm{F}\exists x A$	$\mathrm{F}A_c^x$

This type of rule may be applied many times to the same signed universal formula by choosing all the constants that have already appeared in the tableau.

- By an *existential type* rule a wf A may be instantiated with any constant which *has not occurred previously* in the tableau.

α	$\alpha(c)$
$\mathrm{F}\forall x A$	$\mathrm{F}A_c^x$
$\mathrm{T}\exists x A$	$\mathrm{T}A_c^x$

This type of rule is only applied once to each signed existential formula.

It is possible to prove that the root of a tableau built up by the above rules is satisfiable if and only if one of the branches is satisfiable. This means that the tableau method is sound as only contradictory formulae are the roots of closed tableaux. Furthermore, the method is complete since it is possible to show that the tableau will close if the root is unsatisfiable. We may therefore conclude that every valid sentence A will have an associated closed tableau with $\mathrm{F}A$ as the root.

For $\mathcal{A} = \{A_1, A_2, \ldots\}$, let $\mathrm{T}[S_i]\mathcal{A}$ stand for the set $\{\mathrm{T}[S_i]A_1, \mathrm{T}[S_i]A_2, \ldots\}$ of signed formulae and similarly for $\mathrm{F}[S_i]\mathcal{A}$. Let us now consider the external language L^B and let Γ be a set of sentences and A and B sentences of the language. The rules as given above are used with the addition of the socalled *attachment rule*, namely:

- A branch with the signed formulae $\mathrm{T}[S_i]\Gamma$ and $\mathrm{F}[S_i]A$ closes if $\Gamma \vdash_{\rho(i)} A$.

The attachment rule is sound because if $\Gamma \vdash_{\rho(i)} A$, then, by lemma 4.1, any branch containing the formulae $\mathrm{T}[S_i]\Gamma$ and $\mathrm{F}[S_i]A$ will not be satisfiable (i.e. will close), so the set $[S_i]\Gamma$ entails $[S_i]A$. It can be shown that these rules are also complete. Informally the argument is as follows: Consider any branch containing $\mathrm{T}[S_i]\Gamma$ and $\mathrm{F}[S_i]A$. By lemma 4.1 the set $\mathrm{T}[S_i]\Gamma \cup \{\mathrm{F}[S_i]A\}$ can only be unsatisfiable if $\Gamma \vdash_{\rho(i)} A$. Suppose the branch is open. Then no such derivation can exist otherwise the branch would have been closed by application of the attachment rule above. Thus the formulae are satisfiable.

As an example of the use of the attachment rule we now prove the following, assuming that agent S_i has a saturated deduction structure:

$[S_i](A \to B) \to ([S_i]A \to [S_i]B)$.

(This was already proved earlier in the chapter, using semantic arguments.)

(1)	$\mathrm{F}[S_i](A \to B) \to ([S_i]A \to [S_i]B)$	
(2)	$\mathrm{T}[S_i](A \to B)$	(1)
(3)	$\mathrm{F}([S_i]A \to [S_i]B)$	(1)
(4)	$\mathrm{T}[S_i]A$	(3)
(5)	$\mathrm{F}[S_i]B$	(3)
	$\times$	
	$A, A \to B \vdash_{\rho_i} B$	(2,4,5)

The attachment rule was used to close the table. Because agent S_i is an ideal reasoner modus ponens is one of his deduction rules or is derivable from his deduction rules.

Suppose the signed formula $\mathrm{F}[S_i]A$ is added to a tableau under construction. We know the branch will close if we are able to show that A is a belief. In order to try to show that A is a belief we then start an *auxiliary tableau* with $\mathrm{F}A$ as root. The main tableau represents the outside observer's view of the agent while the auxiliary tableau represents the internal proof process of agent i. If n formulae $\mathrm{F}[S_i]A_j$, $1 \leq j \leq n$, are added to the main tableau n auxiliary tableaux can be formed with respective roots $\mathrm{F}A_j$. (When there is more than one agent the tableaux must be indexed.)

Now suppose at some point we add a signed formula $\mathrm{T}[S_i]B$ to a branch of a main tableau which has one or more auxiliary tableaux for agent i. Then $\mathrm{T}B$ can be added to all branches of each of these auxiliary tableaux because B must be one of the agent's beliefs. If we succeed in closing an auxiliary tableau it means that the belief sentences which we have added to that tableau are inconsistent with the agent's rules and hence the main tableau branch from which they have come must also close.

As a first example let us again prove the following where agent i is an ideal reasoner:

$[S_i](A \to B) \to ([S_i]A \to [S_i]B)$.

(1)	$\mathrm{F}[S_i](A \to B) \to ([S_i]A \to [S_i]B)$	
(2)	$\mathrm{T}[S_i](A \to B)$	(1)
(3)	$\mathrm{F}([S_i]A \to [S_i]B)$	(1)
(4)	$\mathrm{T}[S_i]A$	(3)
(5)	$\mathrm{F}[S_i]B$	(3)
	$\times$	

(6)	$\mathrm{F}B$	(5)
(7)	$\mathrm{T}A$	(4)
(8)	$\mathrm{T}A \to B$	(2)

(9)	$\mathrm{F}A$	(8)	(10)	$\mathrm{T}B$	(8)
	$\times$	(7,9)		$\times$	(6,10)

The auxiliary tableau headed by $\mathrm{F}B$ in line (6) was constructed after the signed formula in line (5) namely $\mathrm{F}[S_i]B$ was added to the main tableau. Line (7) originates from line 4 and line (8) from (2). Because both the branches of the auxiliary tableau close the main tableau closes. The purpose of the auxiliary tableau is to do the derivation $A, A \to B \vdash_{\rho_i} B$.

As a second example we show that, if agent S_i believes that B follows from A, then, if he does not believe B he will not believe A:

(1)	$\mathrm{F}[S_i](A \to B) \to (\neg[S_i]B \to \neg[S_i]A)$	
(2)	$\mathrm{T}[S_i](A \to B)$	(1)
(3)	$\mathrm{F}\neg[S_i]B \to \neg[S_i]A$	(1)
(4)	$\mathrm{T}\neg[S_i]B$	(3)
(5)	$\mathrm{F}\neg[S_i]A$	(3)
(6)	$\mathrm{F}[S_i]B$	(4)
(7)	$\mathrm{T}[S_i]A$	(5)
	$\times$	

(8)	$\mathrm{F}B$	(6)
(9)	$\mathrm{T}A$	(7)
(10)	$\mathrm{T}A \to B$	(2)

(11)	$\mathrm{F}A$	(10)	(12)	$\mathrm{T}B$	(10)
	$\times$	(9,11)		$\times$	(8,12)

The auxiliary tableau starts in line (8) and splits into two branches which both close.

As a final example suppose we have two agents. Agent S_1 has two base beliefs namely $\neg A$ and $S_2 A \to A$ (i.e. 'If agent S_2 believes the sentence A then A is true'). Suppose agent S_2 has only one base belief namely A and neither agent has any deduction rules. We assert agent S_1 will believe $A \wedge \neg A$ and show it as follows:

(1)	$\mathrm{F}[S_1](A \wedge \neg A)$	
(2)	$\mathrm{T}[S_1]\neg A$	given
(3)	$\mathrm{T}[S_1]([S_2]A \rightarrow A)$	given
(4)	$\mathrm{T}[S_2]A$	given
	$\times$	

(5) $\mathrm{F}A \wedge \neg A$ (1)

(6)	$\mathrm{F}A$	(5)				(11)	$\mathrm{F}\neg A$	(5)
(7)	$\mathrm{T}[S_2]A \rightarrow A$	(3)				(12)	$\mathrm{T}\neg A$	(2)
(8)	$\mathrm{F}[S_2]A$	(7)	(10)	$\mathrm{T}A$	(7)		$\times$	(11,12)
(9)	$\times$	(4,8)		$\times$	(6,10)			

The auxiliary tableau starting in line (5) closes because all branches originating from it close. Hence the main tableau is also closed and we have succeeded in showing that agent S_1 believes the contradiction $A \wedge \neg A$. Note that this does not mean that he believes all sentences of the internal language since his deduction structure is not logically complete.

Is it possible to use this proof method when we have sentences of the external language L^{Bq}? Yes. All that is required is to replace the attachment rule for the language L^B given above by a different attachment rule, namely:

- A branch with the signed formulae $\mathrm{T}[S_i]\Gamma$ and $\mathrm{F}[S_i]A$ closes if $\Gamma^\bullet \vdash_{\rho(i)} A^\bullet$.

As an example of the use of the (new) attachment rule let us show that the following holds:

$\exists x_i[S_j]P_1^1(x_i) \rightarrow [S_j]\exists x_i P_1^1(x_i)$

where we assume that the deduction rules of agent S_j are such that he is capable of existential generalisation. (This was already proved earlier in the chapter, using a semantic argument.)

(1)	$\mathrm{F}\exists x_i[S_j]P_1^1(x_i) \rightarrow [S_j]\exists x_i P_1^1(x_i)$	
(2)	$\mathrm{T}\exists x_i[S_j]P_1^1(x_i)$	(1)
(3)	$\mathrm{F}[S_j]\exists x_i P_1^1(x_i)$	(1)
(4)	$\mathrm{T}P_1^1(\bullet a_k)$	(2)
	$\times$	
	$P_1^1(c_k) \vdash_{\rho_j} \exists x_i P_1^1(x_i)$	

The bullet constant $\bullet a_k$ in line (4) is replaced by a schematic constant c_k by the bullet deletion transform of the sentence.

4.8 Summary

The main advantage of Konolige's approach is that resource-bounded agents can be described. We have seen that no restriction need be placed on the deduction rules except that the rules should be applied exhaustively, i.e. that the set of beliefs should be deductively closed. Furthermore, a predicate language is used where quantifying-in is possible and a sound and complete proof theory is given. Other approaches to the problem of logical omniscience exist, for example semantics involving impossible or nonstandard worlds. In

[Fagin et al 1995a] the authors consider the implications of basing a logic of knowledge on a nonstandard logic rather than on standard propositional logic thereby (hopefully) alleviating the logical omniscience problem by an appropriate choice of the nonstandard logic.

Konolige's approach shares the following omission with the approaches of Moore and Levesque: belief sets of agents are considered to be static since no attention is given to changes in their beliefs, in other words belief sets are never revised. In the following chapter we will take a brief look at belief revision.

Chapter 5

Belief Revision

What I know now is only partial; then it will be complete.
1 Cor 13:12

In the preceeding chapters we looked at different ways to describe the beliefs of an agent given some initial beliefs. We did not consider the possibility of information that changes as new facts become available. In other words we did not consider ways to alter the set of beliefs of an agent, either by discarding some of the previous beliefs or by adding new beliefs to the set. In this chapter we take a brief look at the revision of belief sets where a belief set will be understood to be a set of (nonmodal) propositions closed under entailment. Introspective beliefs are specifically excluded, hence a nonmodal language suffices.

Let $\mathcal{K}$ be the belief set or knowledge base of an agent and suppose new information becomes available. If the new information is consistent with $\mathcal{K}$ the set can simply be extended with it. If, however, the new information is inconsistent with $\mathcal{K}$ decisions must be made on how to handle the situation. Say, for example, the new information casts doubt on parts of the knowledge base, i.e. contradicts some wfs in $\mathcal{K}$. Then some kind of revision of $\mathcal{K}$ becomes necessary. Some of the members of the belief set will have to be removed. Which? Well, the belief set was not easily constructed, so we do not want to discard the whole set. We have to decide which of the *reasons* for the inconsistency must be thrown out and which of the *consequences* must be removed.

The problem is that $\mathcal{K}$ is not a simple set of atomic facts: there are complex logical dependencies between the sentences. Logical considerations alone will not be enough to decide which wfs have to be removed. Various principles have been proposed to describe how a belief set should be revised. We discuss the Alchourrón-Gärdenfors-Makinson (AGM) theory ([Blackburn et al 1997]).

5.1 Kinds of Belief Revision

Let $\mathcal{F}$ be the set of all wfs of some propositional language. A belief set is now formally defined as follows:

- A *belief set* $\mathcal{K} \subseteq \mathcal{F}$ is a set which is closed under entailment, i.e. $\mathrm{Cn}(\mathcal{K}) \subseteq \mathcal{K}$.

A sentence A is *accepted* by $\mathcal{K}$ if $A \in \mathcal{K}$ and *rejected* by $\mathcal{K}$ if $\neg A \in \mathcal{K}$.

There seems to be two basic kinds of belief revision, namely (i) to insert or accept information (i.e. adding sentences to $\mathcal{K}$) or (ii) to delete information. The question is how

and when to perform these basic operations. It can be done in a direct or an indirect mode. In the direct mode information is inserted or deleted without bothering about consistency, but in order to determine which conclusions can actually be drawn a complex set of inference rules must be provided. So the complexity of belief revision is shifted inside the inference 'engine'. We will, however, follow the indirect approach.

In the indirect approach one tries to perform belief revision according to (at least some of) the following guidelines:

- *Consistency.* The sentences of a belief set should be kept consistent whenever possible.
- *Closure.* If a sentence A is entailed by the belief set, then A should be a member of the set.
- *Minimality.* The amount of information lost when the belief set is revised should be kept to a minimum.
- *First Things Last.* If some beliefs are in some way considered as more important than others, the least important beliefs should be discarded first.

By following these guidelines belief revision becomes highly nontrivial but the advantage is that standard logic can then be used as the underlying 'inference engine'.

The AGM theory is the most prominent example of belief revision in the indirect mode. In this approach three main kinds of belief revision are considered. We list them below and illustrate each using the language with three atoms p, q and r. We think of p as 'I have an elder brother', of q as 'I have an elder sister' and of r as 'I have a younger sister'. Suppose now our agent does not, as far as he knows, have an elder brother and has one sister (older than him). We think of the agent's belief set as the set of all consequences of his initial beliefs, thus $\mathcal{K} = \mathrm{Cn}(\{\neg p, q\})$.

- *Expansion.* Suppose new information becomes available in the form of the sentence A and suppose A is consistent with the belief set $\mathcal{K}$. Then $\mathcal{K}$ is *expanded with* A, denoted by $\mathcal{K} + A$, formally defined as:

  $\mathcal{K} + A = \mathrm{Cn}(\mathcal{K} \cup \{A\})$.

  The expanded set will contain $\mathcal{K}$ as a subset.

  In our example, suppose a baby sister is born, i.e. the fact r becomes available. Then we will have $\mathcal{K} + A = \mathrm{Cn}(\{\neg p, q, r\})$.

- *Revision.* Suppose new information becomes available in the form of the sentence A and suppose A is inconsistent with the belief set $\mathcal{K}$. If we simply expand $\mathcal{K}$ with A the resultant belief set will be the set of all wfs. Thus expansion is not appropriate. Some other operation must be defined so that consistency will be preserved. This means some sentences of $\mathcal{K}$ will have to be discarded if A is included. We say $\mathcal{K}$ is *revised with* A and denote the resultant belief set by $\mathcal{K} * A$. It is possible that this resultant set is neither a superset nor a subset of $\mathcal{K}$. We will not propose any specific revision operation but will look at general properties that such operations should satisfy.

  Suppose in our example the fact $A = p \vee \neg q$ becomes available. To include this new information we will have to revise $\mathcal{K}$ and either $\neg p$ or q will not be a member of the revised set. There is, however, no purely logical reason for choosing which information should be discarded.

- *Contraction.* A belief set $\mathcal{K}$ is *contracted with* A when a sentence $A \in \mathcal{K}$ is retracted from $\mathcal{K}$ without new information being added. Because of the definition of belief sets it will, in general, be necessary to discard further sentences of $\mathcal{K}$ in order to retain closure under entailment. The resultant belief set is denoted by $\mathcal{K} - A$ and it will be a subset of $\mathcal{K}$. As in the case of revision we do not propose any specific contraction operation but give the AGM postulates that such operations should satisfy.

  In our example the agent might wonder how his life would have been if he did not have an elder sister ...

5.2 Postulates for Revision

One should note that the following two characteristics are important features of the AGM approach:

- *Minimal Change.* As much as possible of the original belief set should be retained.
- *Functionality.* We assume the operation * is a function from belief sets and sentences to belief sets, i.e. we assume that for every belief set $\mathcal{K}$ and sentence A, there is a unique belief set $\mathcal{K} * A$.

The first rule, called the *closure* postulate, states that the result of revision should again be a belief set:

- For any wf A and any belief set $\mathcal{K}$, we have that $\mathcal{K} * A$ is a belief set. **(*1)**

The second rule, called the *success* postulate, states that incoming information has priority over old information:

- $A \in \mathcal{K} * A$. **(*2)**

The third and fourth postulates are called the *expansion* postulates. The third postulate ensures that revising with A will never result in more information than expanding with A:

- $(\mathcal{K} * A) \subseteq (\mathcal{K} + A)$. **(*3)**

The fourth postulate is the following:

- If $\neg A \notin \mathcal{K}$, then $(\mathcal{K} + A) \subseteq (\mathcal{K} * A)$. **(*4)**

The fourth postulate together with the third state that, if A is consistent with the (original) belief set, then revising with A and expanding with A yield the same result. This is in accordance to the minimality principle. We give it as a remark:

Remark 5.1 *If* $\neg A \notin \mathcal{K}$, *then* $(\mathcal{K} + A) = (\mathcal{K} * A)$.
The result follows immediately from postulates (*3) and (*4). ♠

What will the result be if we revise a belief set with a tautology, i.e. with a wf which is valid in all frames? Well, any belief set has all the tautologies as members, in other words revising by any of them will not change the set:

Remark 5.2 *Given any consistent belief set $\mathcal{K}$ and any tautology A, it follows that $\mathcal{K} * A = \mathcal{K}$.*

The result follows immediately. ♠

As noted above a belief set should be consistent if at all possible. So $\mathcal{K} * A$ should be consistent whenever A is. This is the purpose of the fifth rule, called the *consistency preservation* postulate:

- $\mathcal{K} * A = \mathcal{F}$ iff $\models \neg A$. **(*5)**

Thus the fifth postulate requires that the result of revising with A should always be consistent except when $\neg A$ is a tautology, in other words $\mathcal{K} * A$ should be consistent except when A is a contradiction (i.e. false in all worlds).

The sixth postulate, the *extensionality* postulate, ensures that the syntactic form of A does not have any effect when revision is done with it:

- If $\models A \leftrightarrow B$, then $\mathcal{K} * A = \mathcal{K} * B$. **(*6)**

The converse of the sixth postulate does not always hold. We can easily see it by considering the case $\mathcal{K} = \mathrm{Cn}(\{p, q\})$ and $A = p$ and $B = q$.

The above six postulates are the basic rules for the way in which a belief set should be revised. We will shortly give two additional ones, but first have a look at some consequences of these six postulates.

Firstly, revision is not 'commutative', thus in general $(\mathcal{K} * A) * B \neq (\mathcal{K} * B) * A$. (Take, for example, $\mathcal{K} = \mathrm{Cn}(\{p\})$, $A = q$ and $B = \neg q$.) This means that the order in which revision is done is important. Furthermore it is generally not the case that $(\mathrm{Cn}(A)) * B = (\mathrm{Cn}(B)) * A$. (Again we can see it by taking $A = q$ and $B = \neg q$.)

Secondly, recall that the Cn operator is monotonic, i.e. given two sets Φ and Γ, if $\Phi \subseteq \Gamma$, then $\mathrm{Cn}(\Phi) \subseteq \mathrm{Cn}(\Gamma)$. Then the fourth postulate above is equivalent to stating that $\mathcal{K}$ is a subset of $\mathcal{K} * A$ if A is consistent with $\mathcal{K}$. It can be shown as follows:

Remark 5.3 *The following two rules are equivalent:*

(i) **(*4)** *If A is consistent with $\mathcal{K}$, then $(\mathcal{K} + A) \subseteq (\mathcal{K} * A)$*

and

(ii) if A is consistent with $\mathcal{K}$, then $\mathcal{K} \subseteq \mathcal{K} * A$.

First assume *(ii)* and that A is consistent with $\mathcal{K}$, i.e. $\neg A \notin \mathcal{K}$. Then

$\mathcal{K} + A$	$= \mathrm{Cn}(\mathcal{K} \cup \{A\})$	Definition of expansion
	$\subseteq \mathrm{Cn}((\mathcal{K} * A) \cup \{A\})$	By *(ii)* and the monotonicity of Cn
	$\subseteq \mathrm{Cn}(\mathcal{K} * A)$	Postulate (*2)
	$= \mathcal{K} * A$	Postulate (*1)

Secondly, assume that the fourth postulate, i.e. *(i)*, holds and that A is consistent with $\mathcal{K}$, i.e. $\neg A \notin \mathcal{K}$. Then

$\mathcal{K}$	$\subseteq \mathrm{Cn}(\mathcal{K})$	Definition of Cn
	$\subseteq \mathrm{Cn}(\mathcal{K} + A)$	Monotonicity of Cn
	$= \mathcal{K} + A$	$\mathcal{K} + A$ is a belief set
	$\subseteq \mathcal{K} * A$	Postulate (*4) ♠

The two additional postulates for revision involve revising a belief set with a conjunction. They are called the *conjunction* postulates. It seems reasonable to expect the result of revising a belief set $\mathcal{K}$ with a conjunction $A \wedge B$ to be equivalent to revising with A and then revising with B. This turns out to be the case provided that B is consistent with $\mathcal{K} * A$ (in which case, of course, $(\mathcal{K} * A) * B = (\mathcal{K} * A) + B$). It is customary to split the result into two parts. So we get the following two postulates:

- $\mathcal{K} * (A \wedge B) \subseteq (\mathcal{K} * A) + B$. **(*7)**
- $\neg B \notin \mathcal{K} * A$, then $(\mathcal{K} * A) + B \subseteq \mathcal{K} * (A \wedge B)$. **(*8)**

Postulates (*7) and (*8) are very powerful. Some of the other postulates become derived rules in their presence. We will show two cases:

Remark 5.4 *Postulate (*3) is a special case of postulate (*7), assuming postulate (*6) holds.*

Let $\mathcal{K}$ be a belief set, A any wf and B a tautology. Then we have

$$\begin{array}{lll} \mathcal{K} * A & = \mathcal{K} * (B \wedge A) & \text{Postulate (*6)} \\ & \subseteq (\mathcal{K} * B) + A & \text{Postulate (*7)} \\ & = \mathcal{K} + A & \text{Remark 5.2} \end{array}$$ ♠

Remark 5.5 *Postulate (*4) is a special case of postulate (*8), assuming postulate (*6) holds.*

Let $\mathcal{K}$ be a belief set, A any wf such that $\neg A \notin \mathcal{K}$ and B a tautology. Then we have

$$\begin{array}{lll} \mathcal{K} + A & = (\mathcal{K} * B) + A & \text{Remark 5.2} \\ & \subseteq \mathcal{K} * (B \wedge A) & \text{Postulate (*8)} \\ & = \mathcal{K} * A & \text{Postulate (*6)} \end{array}$$ ♠

5.3 Postulates for Contraction

The idea of contraction with a wf A is to have a belief set which no longer contains A if that is at all possible. The very first requirement is that the result of a contraction of a belief set should again be a belief set. This is called the *closure* postulate.

- For any wf A and any belief set $\mathcal{K}$, we have that $\mathcal{K} - A$ is a belief set. **(-1)**

The second rule, the *inclusion* postulate, states that the resultant set should be a subset of the original one:

- For any wf A and any belief set $\mathcal{K}$, we have that $\mathcal{K} - A \subseteq \mathcal{K}$. **(-2)**

Suppose we contract with a wf A but $A \notin \mathcal{K}$. Then the belief set should be unchanged. This is stated by the third postulate, namely the *vacuity* postulate:

- If $A \notin \mathcal{K}$, then $\mathcal{K} - A = \mathcal{K}$. **(-3)**

Suppose the wf A with which we contract is a tautology. Then, since belief sets are closed under entailment, it will still be in the belief set after contraction. In all other cases, though, A should not be a member of the resultant belief set. This is the fourth rule, called the *success* postulate:

- If $\not\models A$, then $A \notin \mathcal{K} - A$. (-4)

If $A \in \mathcal{K}$, then it follows from the above four postulates and the monotonicity of Cn that $(\mathcal{K} - A) + A \subseteq \mathcal{K}$ whenever $A \in \mathcal{K}$. Can this set inclusion be replaced by an equality? In other words, if an agent believes A, is it the case that consecutively removing it from the belief set and then adding it again will result in the original belief set? This is what the fifth postulate states: enough must be left of the belief set when removing A from it so that it can be restored by expanding with A again. This postulate is called the *recovery* postulate.

- If $A \in \mathcal{K}$, then $(\mathcal{K} - A) + A = \mathcal{K}$. (-5)

There are, however, situations where one would not want the recovery postulate to hold. As an example we give one cited by Hansson ([Hansson 1996]). Suppose we have the language with two atoms p and q where we think of p as 'Cleopatra had a daughter' and of q as 'Cleopatra had a son'. Let the belief set $\mathcal{K}$ of the agent be $\mathrm{Cn}(p \wedge q)$, i.e. the agent believes Cleopatra had both a son and a daughter, information he got when he read a book called 'Life and times of Cleopatra'. Both p and q will be members of the set $\mathcal{K}$. Now suppose a friend of the agent informs him that 'Life and times of Cleopatra' is a novel, not based on historical fact. The agent now contracts his belief set with $p \vee q$, i.e. he no longer believes that Cleopatra had any child. Suppose after that, however, he learns from a (reliable) encyclopedia that Cleopatra actually had a child. Then it would be reasonable for the agent to expand his belief set with $p \vee q$ but *without reintroducing either* p *or* q. Thus $(\mathcal{K} - p \vee q) + p \vee q \neq \mathcal{K}$ since p and q are members of the original belief set $\mathcal{K}$ but not of the resultant belief set.

As in the case of revision we expect that the syntax of a wf will have no effect, i.e. that contracting with equivalent sentences will result in the same belief set. This is what the *extensionality* postulate, the sixth rule, is all about:

- If $\models A \leftrightarrow B$, then $\mathcal{K} - A = \mathcal{K} - B$. (-6)

One might expect that contraction with a logically 'weaker' wf would lead to smaller belief sets than contraction by 'stronger' wfs, in other words one might expect that, if $A \models B$, then $\mathcal{K} - B \subseteq \mathcal{K} - A$. (The reason for this intuition is that B is 'easier to deduce', so more wfs have to be removed in order not to be able to deduce B.) This is, however, *not* the case. To see this, let us look at an example, using the language with three atoms, p, q and r. We think of p as 'James plays rugby', of q as 'Louis plays rugby' and of r as 'Louis makes lots of money'. Suppose the agent believes that both James and Louis are rugby players and that, if Louis plays rugby, then he makes lots of money. So his belief set is $\mathcal{K} = \mathrm{Cn}(\{p, q, q \rightarrow r\})$. This set will include the wf r. If the agent no longer believes that James is a rugby player we intuitively feel that r should still be in his belief set. If he, however, gives up his belief that both James and Louis are rugby players, r should no longer be in his belief set. (He may still have $p \vee q$ in his belief set but if he is no longer certain that q, he should not have r as a belief.) So $r \in \mathcal{K} - p$, but $r \notin \mathcal{K} - (p \wedge q)$. Thus, contrary to our intuition, we have $\mathcal{K} - p \not\subseteq \mathcal{K} - (p \wedge q)$.

An unwelcome result is that monotonicity does not hold for contraction. It is thus *not* the case that, if $\mathcal{K}' \subseteq \mathcal{K}$, then $\mathcal{K}' - A \subseteq \mathcal{K} - A$. We can illustrate this by considering two belief sets, $\mathcal{K}' = \mathrm{Cn}(p \vee q)$ and $\mathcal{K} = \mathrm{Cn}(p)$. We have that $\mathcal{K}' \subseteq \mathcal{K}$. Suppose we contract both belief sets by p. Then $\mathcal{K}'$ does not change but $\mathcal{K} - p$ will be different from $\mathcal{K}$. It may be the

case that $\mathcal{K} - p$ contains no sentences except the tautologies, hence $\mathcal{K}' - p \not\subseteq \mathcal{K} - p$. The absence of monotonicity is an indication that definitions of contraction are much harder than definitions of expansion. We do have, however, rather neat postulates on contractions with conjunctions.

The rugby player example above has shown that it is in general not the case that $(\mathcal{K} - A) \subseteq \mathcal{K} - (A \wedge B)$, but the seventh postulate, one of the *conjunction* postulates, states a related rule:

- $(\mathcal{K} - A) \cap (\mathcal{K} - B) \subseteq \mathcal{K} - (A \wedge B)$. **(-7)**

Is it maybe the case that $\mathcal{K} - (A \wedge B) \subseteq (\mathcal{K} - A)$? It holds only when $A \notin \mathcal{K} - (A \wedge B)$. This is the eighth (the other *conjunction*) postulate:

- $A \notin \mathcal{K} - (A \wedge B)$, then $\mathcal{K} - (A \wedge B) \subseteq (\mathcal{K} - A)$. **(-8)**

It is possible to prove that, assuming that postulates (-1) - (-6) hold, the conjunction postulates (-7) and (-8) hold iff

$$\mathcal{K} - (A \wedge B) = \begin{cases} \mathcal{K} - A, & \text{or} \\ \mathcal{K} - B, & \text{or} \\ (\mathcal{K} - A) \cap (\mathcal{K} - B). \end{cases}$$

5.4 Revisions and Contractions

In any system where two operations are defined the question arises how these operations interact. We now consider two postulates which combine revision and contraction.

In order to revise a belief set $\mathcal{K}$ with a wf A it is necessary to remove all wfs $B \in \mathcal{K}$ which contradict A, in other words a contraction needs to be done before A can be added. The main reason for a contradiction will be the presence of $\neg A$. It therefore seems reasonable to regard revision with A as contraction with $\neg A$, i.e. get $\mathcal{K} - \neg A$, followed by expansion with A. This is expressed by the *Levi identity* ([Blackburn et al 1997]):

- $\mathcal{K} * A = (\mathcal{K} - \neg A) + A$.

Suppose we have a contraction operation which satisfies the postulates given in the previous section and suppose we define revision by the Levi identity. Will it be the case that the revision postulates are preserved? Fortunately the answer is positive. If the six postulates (-1) to (-6) hold for a contraction function -, then the revision function defined by the Levi identity satisfies (*1) to (*6). Also, if the seven postulates (-1) to (-7) hold for a contraction function -, then the revision function defined by the Levi identity also satisfies (*7) and, if the seven postulates (-1) to (-6) and (-8) hold for a contraction function -, then the revision function defined by the Levi identity also satisfies (*8).

It is also possible to give a definition of contraction in terms of revision, namely by the *Harper identity* ([Blackburn et al 1997]):

- $\mathcal{K} - A = \mathcal{K} \cap (\mathcal{K} * \neg A)$.

[Levesque 1990] LEVESQUE H J. All I know: a study in autoepistemic logic. *Artificial Intelligence*. Vol 42. 1990.

[Lukaszewicz 1990] LUKASZEWICZ W. Non-monotonic reasoning: formalization of commonsense reasoning. Ellis Horwood Limited. 1990.

[McDermott et al 1987] MCDERMOTT D, DOYLE J. Non-monotonic logic I. In *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann Publishers, Inc. 1987.

[Moore 1984] MOORE R C. Possible-world semantics for autoepistemic logic. *Proceedings 1984 Non-monotonic Reasoning Workshop*, New Paltz, NY. 1984.

[Moore 1985] MOORE R C. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*. Vol 25. 1985.

[Shoenfield 1967] SHOENFIELD J R. Mathematical logic. Addison-Wesley Publishing Company. 1967.

[Stalnaker 1980] STALNAKER R C. A note on nonmonotonic modal logic. Department of Philosophy, Cornell University, Ithaca, NY. 1980.

Index