Resource-Bounded Reasoning about Knowledge

Dipl.Inf. **Ho Ngoc Duc**

April 2001

Acknowledgments

First of all, I wish to thank my supervisor Heinrich Herre for his excellent guidance throughout my studies at the University of Leipzig.

I am very much indebted to Peter Steinacker who introduced me to logic and who has always supported and encouraged me during the past several years.

Parts of this thesis were done at the Australian National University, Canberra. I thank John Slaney and other members of the Automated Reasoning Project at the ANU for their helpful discussions.

I thank the numerous other people with whom I had the chance to discuss the various problems I try to solve in the thesis. Their constructive comments have helped me very much to clarify my thoughts and to organize the thesis better. I am especially grateful to Jens Dietrich, Siegfried Gottwald, Georg Meggle, Gerd Wagner, Heinrich Wansing, and Georg Henrik von Wright.

I dedicate this work to my parents.

Contents

Introduction

1.1 Motivation

The formal analysis of reasoning about knowledge has attracted much attention recently. Epistemic logic was invented in the early 1960's by philosophers as a tool for describing epistemic concepts such as knowledge and belief formally.[1] At the beginning, the main interest was to find inherent properties of knowledge (and related concepts) and to apply the analysis to epistemology. More recently, researchers from other disciplines such as linguistics, economics, game theory, and computer science have become increasingly interested in reasoning about knowledge. In addition to the more traditional topics, many other questions have become relevant for those who are more interested in applications, e.g., questions about computational complexities or the relationship between and agent's knowledge and his action.

Within computer science, reasoning about knowledge plays an extremely important role in contemporary theories of intelligent agents. In recent years a number of approaches have been proposed in (Distributed) Artificial Intelligence (DAI) to specify rational agents in terms of mental qualities like knowledge, belief, want, goal, commitment, and intention. There is no universally accepted definition of the term "agent" in the literature, yet there seem to be a common picture of artificial agents within the DAI community: "agents" are, or should be, formal versions of human agents, possessing formal versions of mental attitudes like knowledge, belief, goals. In short, the notion of an "intentional stance" ([Den87], [McC79]) is adopted. It has proved possible and *useful* to characterize agents using those attitudes.[2] There is no clear consensus in the DAI community about precisely which combination of mental attitudes is best suited to characterizing agents. However, it seems to be an agreement that belief (or knowledge) should be taken as one of the basic notions of the agent theory ([WJ95]).

The emphasis on epistemic concepts is not accidental. First, the role that knowledge plays in decision and action is obvious. Second, knowledge and belief are most intensively studied among all mentalistic concepts. In fact, the other concepts are usually modeled after the way the epistemic ones are modeled. Third, epistemic concepts are arguably among the most fundamental mental notions: many other mentalistic

[1] Sometimes the term "epistemic logic" is reserved for the logic of knowledge, and "doxastic logic" is used to denote that of belief. In the present thesis we shall use the term "epistemic logic" in the wider sense.

[2] See [FG97] for a recent discussion of the agent concept. McCarthy ([McC79]) discusses the problem of ascribing human-like qualities to artificial entities.

concepts seem to be derivable from the epistemic ones, but not vice versa. For example, an old philosophical thesis states that the concept of desire is reducible to that of belief: an agent desires something if he believes that having it is useful. A discussion of this desire-as-belief thesis can be found in [Lew88], [Lew96]. The normative concepts of obligation and permission could also be reduced to the concept of belief. Anderson's reduction of deontic logic to alethic modal logic ("something is obligatory if and only if not doing it necessarily leads to punishment", cf. [And58], [And67]) can be interpreted epistemically as: "something is obligatory if and only if the agent knows (or believes) that not doing it leads to punishment". It can be shown that under this epistemic interpretation, the deontic axioms can be derived from axioms of epistemic logic.

In short, formal theories of knowledge constitute the most important foundation for theories of agency. Consequently, all strengths and weaknesses of the underlying epistemic theory propagate to the agent theory based on it. We will see that this has important consequences for the suitability of agent theories for characterizing intelligent agents.

Typically, formal theories of agents are used as internal specification languages, i.e., languages used by agents to reason about themselves and about other agents. As such, agent theories must describe agents accurately and realistically. In order to interact with each other, each agent needs an accurate representation of themselves and of other agents, their information states, their preferences et cetera. I shall show that this requirement cannot be met if mainstream epistemic logic is used to model an agent's cognitive state.

The purpose of my thesis is to provide a more suitable epistemic foundation to theories of intelligent agents. I will argue that agent theories need to be based on better logics of knowledge than the ones on which they are based now. The main reason is that agents — both human and non-human — are inherently resource-bounded: they cannot perform arbitrarily complex reasoning tasks within constant, limited time. Mainstream modal epistemic logic, however, is not able to account for that resource boundedness. The most obvious indication of this inability is the so-called logical omniscience problem of epistemic logic. I shall show that almost all work that purports to be about knowledge is done under assumptions that are unreasonable for knowledge of realistic, resource-bounded agents. Then I will propose some systems of epistemic logic which can be used for resource-bounded reasoning[3].

1.2 Structure of the thesis

The thesis is organized as follows. First, I shall review briefly the possible-worlds approach to epistemic logic and its relationship to recent agent theories in chapter 2. I shall show that the modal approach can at most account for the concept of implicit knowledge, but that concept is not helpful in describing agents, simply because agents need to act upon what they *explicitly* know, and not what they merely *implicitly* know.

In chapter 3 I examine some influential alternative approaches to epistemic logic

[3]Some authors use the term "bounded rationality" to express the idea that an agent cannot compute everything he could if his resources were unlimited. That term is somewhat misleading, so I shall use "resource boundedness" throughout the thesis.

and evaluate how they may be suited to describing realistic agents. I shall show that almost all attempts proposed in the literature to overcome the difficulties of the modal approach consist in weakening the standard epistemic systems. That is, weaker systems are considered where the agents do not possess the full reasoning capacities of ideal reasoners. I shall argue that those alternatives are not satisfactory because they restrict the agents' intelligence arbitrarily, so the intuition that agents *are* rational gets lost. Consequently, they are not suitable for formalizing the informational aspect of intelligent agents.

In chapter 4 a framework for reasoning about explicit knowledge will be developed. I shall show that axioms for explicit knowledge must have the following form: if the agent knows all premises of a valid inference rule, and if he thinks hard enough, then he will know the conclusion. To formalize such an idea, I propose to introduce a dynamic component into the epistemic language. I shall show that my approach offers an intuitive solution to the logical omniscience problem while preserving the intuition that agents are rational. My approach is therefore is suitable for formalizing the notion of actual, or explicit knowledge.

In chapter 5 I shall develop logics of algorithmic knowledge — a new concept of knowledge which generalizes both implicit and explicit knowledge. The main idea is to combine epistemic logic with a complexity analysis: we consider how long an agent will need to compute the solution to a certain problem. After explaining the underlying intuitions I shall introduce the concept of algorithmic knowledge and develop formal theories of this new concept.

A short overview of modal, temporal, and dynamic logic is given in appendix A. Formal proofs of some theorems are found in appendix B.

1.3 Main results

The main result of this thesis is the clarification of some central concepts of agent theories, namely, the concepts describing the informational aspect of intelligent, resource-bounded agents. The main technical result is a framework for establishing direct connections between an agent's knowledge and his available resources. In the thesis two epistemic concepts — the concepts of explicit knowledge and algorithmic knowledge — will be introduced and characterized axiomatically. It will be shown that these concepts are important for resource-bounded reasoning about knowledge and useful for describing rational, but realistic and implementable agents.

Although the thesis is primarily concerned with (Distributed) Artificial Intelligence, I am convinced that it will have a considerable impact on other fields of research, especially on philosophy and game theory.

In the philosophical literature, epistemic logic has been frequently criticized for not being able to model agents realistically. Several researchers have therefore drawn the conclusion that epistemic logic is either not possible, or it is not useful for epistemologists interested in actual knowers in the actual world ([Hoc72], [Hal95]). My attempt to model realistic, resource-bounded reasoners can be seen as a defense against those attacks. By actually specifying a theory of knowledge that can be verified empirically I will provide the evidence that epistemic logic is indeed possible.

In the field of game theory and mathematical economics, resource boundedness has been a primary concern for a long time ([Sim57]). Since Aumann's seminal work ([Aum76]), game theorists have become interested in the role of knowledge (and especially common knowledge) in games. Recent works on the epistemic foundations of games (e.g., [Bin90], [Wal92], [Bac94]) have made clear what implicit assumptions concerning the players' knowledge are made when modeling a game. Because these assumptions are recognized as too strong for realistic agents, several attempts have been made to weaken the underlying epistemic logic in order to describe players more realistically ([Bac94], [LM94], [Hei95]). My investigation can contribute to the search for a more suitable epistemic foundation of game theory.

Some results of this thesis have been published previously. Chapter 4 is based on [Ho95] and [Ho97]. Parts of chapter 5 are based on [Ho98].

1.4 Notations and preliminaries

In the following we shall use the following symbols and abbreviations: ω denotes the set of natural numbers. $=_{def}$ is the symbol for a definition. Pow is the power set function: if X is any set then $Pow(X)$ is the powerset of X. 'wrt' abbreviates 'with respect to', and 'iff' stands for 'if and only if'.

Let X be any set and $R \subseteq X \times X$ a binary relation on X. Then R^+ denotes the transitive closure and R^* the reflexive, transitive closure of R. If R and S are two relations on X then $R;S$ denotes the composition of the two relations, i.e., $(s,t) \in R;S$ iff there exists a u such that $(s,u) \in R$ and $(u,t) \in S$.

Let $R \subseteq X \times X$ a binary relation on a set X. We will be considering relations having certain algebraic properties:

- R is reflexive iff for all $x \in X$, xRx.
- R is transitive iff for all $x,y,z \in X$, if xRy and yRz then xRz.
- R is symmetric iff for all $x,y \in X$, if xRy then yRz.
- R is serial iff for all $x \in X$ there is a $y \in X$ such that xRy.
- R is Euclidean iff for all $x,y,z \in X$, if xRy and xRz then yRz.
- R is directed iff for all $x,y,z \in X$, if xRy and xRz then there is some $t \in X$ such that yRt and zRt.

The relation R is said to be an equivalence relation if it is reflexive, transitive, and symmetric. It is easy to verify that every reflexive relation is also serial and every reflexive, transitive and Euclidean relation is an equivalence relation. Often the term "confluent" is used as a synonym for "directed".

To construct a formal language we will start with a countable set of atomic formulae and use the usual Boolean connectives: negation ($\neg$), conjunction ($\wedge$), disjunction ($\vee$), implication ($\rightarrow$), and material equivalence ($\leftrightarrow$), possibly together with additional (non-extensional) connectives, to form more complex formulae.[4] Atomic formulae will be

[4] We avoid using the word "intensional" because epistemic concepts are not intensional in the sense of Carnap [Car47]. Those concepts are — as Cresswell pointed out ([Cre80] — hyper-intensional.

denoted by $\phi, \phi_0, \phi_1, \ldots$. To denote arbitrary formulae we use $\alpha, \beta, \gamma, \ldots$, possibly with indexes. We take negation and implication as basic connectives. Disjunction, conjunction, and material equivalence are introduced as abbreviations:

$$(\alpha \vee \beta) =_{def} \neg\alpha \rightarrow \beta$$

$$(\alpha \wedge \beta) =_{def} \neg(\neg\alpha \vee \neg\beta)$$

$$(\alpha \leftrightarrow \beta) =_{def} ((\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha))$$

To omit parentheses where possible we also adopt the convention that the binding powers of the connectives decrease in the following order: negation, conjunction, disjunction, implication, and material equivalence.

Let $\mathcal{L}$ be a formal language which contains the above-mentioned Boolean connectives. A logic Λ in this language will be defined by specifying a set of axiom schemata and rules of inference. Formulae derivable (or provable) from the axioms using the inference rules of the logic are called theorems. We often identify a logic with the set of its theorems. We write $\vdash_\Lambda \alpha$ to indicate that α is a Λ-theorem. If $\Gamma \subseteq \mathcal{L}$ and $\alpha \in \mathcal{L}$ then we say that α is Λ-deducible from Γ, denoted $\Gamma \vdash_\Lambda \alpha$, if there exist $\beta_1, \ldots, \beta_n \in \Gamma$ such that $\vdash_\Lambda \beta_1 \wedge \ldots \wedge \beta_n \rightarrow \alpha$. (In the case $n = 0$, this means that $\vdash_\Lambda \alpha$.) We write $\Gamma \nvdash_\Lambda \alpha$ if α is not Λ-deducible from Γ. A set $\Gamma \subseteq \mathcal{L}$ is Λ-consistent if there is no formula α such that $\Gamma \nvdash_\Lambda (\alpha \wedge \neg\alpha)$. The deductive closure of a set $\Gamma \subseteq \mathcal{L}$ with respect to the logic Λ is defined as:

$$Cn_\Lambda(\Gamma) =_{def} \{\alpha \in \mathcal{L} : \Gamma \vdash_\Lambda \alpha\}$$

The semantic counterpart of the provability concept is the concept of validity. We shall define precisely the notion of a model for a formal language and the concept of validity in a model, i.e., we shall specify when a formula $\alpha \in \mathcal{L}$ is valid in some model M (for the language $\mathcal{L}$), in symbol $M \models \alpha$. A formula α is called valid with respect to some class $\mathcal{C}$ of models, denoted $\models_\mathcal{C} \alpha$, if it is valid in all models of that class. A logic is said to be sound with respect to a class of models if and only if all of its theorems are valid wrt that class of models. It is complete wrt a class of models iff all valid formulae of that class of models are theorems of the logic. We say that a logic Λ is determined by a class $\mathcal{C}$ of models just in case it is sound and complete wrt $\mathcal{C}$, i.e., $\vdash_\Lambda \alpha$ iff $\models_\mathcal{C} \alpha$.

Chapter 2

Modal epistemic logic

At the beginning of epistemic logic, attempts were made to develop systems to describe actual knowledge of real agents. The term "knowledge" was originally used in its ordinary language meaning: to say that an agent knows a sentence either means that he consciously assents to it, or that he immediately sees it to be true when the question is presented. However, it was soon realized that describing actual knowledge is a nearly impossible task: actual knowledge does not seem to obey any logic. If we consider real agents and ask what they *actually* know, we can check empirically that an agent's knowledge is often not closed under any logical law. From some epistemic statement one cannot infer reliably any other epistemic statement, i.e., one can hardly find any genuine epistemic statement that may claim universal validity. There seems to be no general epistemic principle that cannot be disproved with a counter-example. It seems impossible to develop a logic of actual knowledge because — to quote Eberle ([Ebe74]) — such a logic must be able to "provide for total ignoramusses (ones who knows nothing), complete idiots (ones who cannot draw even the most elementary inferences), and ultimate fools (ones who believe nothing but contradictions)".

In order to make epistemic logic possible, idealizations were made concerning the reasoning capacities of the agents, and modal systems were proposed to describe such idealized agents. However, the idealizations made by modal epistemic logic are too strong for any realistic agent: they require that agents be very powerful reasoners who know all logical consequences of what they know, including all logical truths. If "knowledge" is interpreted in its normal, ordinary language meaning then such perfectly rational, logically omniscient agents are non-existent. No human agent has the reasoning capacities required by modal epistemic logic. We cannot build artificial agents that possess the reasoning power described by normal modal systems. Thus, modal epistemic logic cannot be interpreted as describing what agents actually know.

To save modal logic as logic of knowledge, a new interpretation of epistemic logic has been proposed: the concept of implicit knowledge is invented, and modal epistemic logic is now interpreted as describing this concept. That is, epistemic logic is not taken as describing what an agent actually knows, but only what is implicitly represented in his information state, i.e., what logically follows from his actual knowledge. What an agent actually knows is called his explicit knowledge.

In the following I review briefly the modal approach to epistemic logic. (An overview of basic modal logic is contained in appendix A.) I shall argue that that approach cannot serve as an adequate foundation for agent theories, because modal epistemic

logic cannot account for the concept of explicit knowledge, but only explicit knowledge can constitute a cognitive state which can play a certain justificatory role for agents' action.

2.1 The "received view": modal epistemic logic

Among all approaches to epistemic logic that have been proposed, the modal approach has been the most widely used for modeling knowledge. An important reason for the popularity of that approach is its simplicity: systems of modal logic are given an epistemic interpretation, and the main technical results about epistemic logic can be obtained almost automatically. To interpret modal logic epistemically one reads modal formulae as epistemic statements expressing the attitude of certain agents towards certain sentences, and the semantics for modal logic is also given a new interpretation.

The interpretation of modal axioms as axioms for knowledge are not without difficulties. If we follow the ordinary usage of the word "knowledge" then that interpretation is certainly wrong. For example, consider the modal formula $\Box\alpha\wedge\Box(\alpha\rightarrow\beta)\rightarrow\Box\beta$. If interpreted epistemically, it says that if an agent knows the two premises of modus ponens then he also knows the conclusion. This is clearly too strong: there may be sentences α and β such that an agent knows both α and $\alpha\rightarrow\beta$ and yet fails to know β. In general, from some epistemic statements one cannot deduce any other epistemic statement. Given the information that an agent's knowledge includes a set Γ of sentences, in reality we can never infer reliably that the agent knows a sentence from the deductive closure $Cn(\Gamma)$ of Γ with respect to a deductive system Cn (except for those already in Γ), even if we suppose Cn to be very weak (but not degenerate in the sense that $Cn(\Gamma)=\Gamma$.) This point has led many people to raise the question if epistemic logic is possible at all, or do we have to leave the realm of logic when reasoning about knowledge and belief ([Hoc72], [Bar89].)

Given the mentioned difficulty, how can we make epistemic logic possible? The answer is idealization. One restricts attention on the class of rational agents, where rationality is defined by certain postulates: agents have to satisfy at least some conditions to qualify as rational. For example, such a condition may read: "If an agent is rational then he should know the laws of logic, therefore, if he knows α and $(\alpha\rightarrow\beta)$, he should be able to use modus ponens to infer β". Those "rationality postulates" for knowledge show a striking similarity with the laws of modal logic, so we may attempt to interpret the necessity operator in modal axioms as knowledge operator and try to justify them as axioms for knowledge. A systematic way to justify epistemic axioms is by way of semantics: one tries to find a plausible epistemic interpretation of a semantics for modal logic. Such an interpretation of the possible worlds semantics was proposed by Hintikka ([Hin62]) and adopted widely hence.

2.1.1 The language of epistemic logic

Suppose that we have a group consisting of N agents. Then we augment the language of propositional logic by N knowledge operators $K_1,\ldots,K_N$ (one for each agent), and

form formulae in the obvious way. A statement like $K_1\alpha$ is read "agent 1 knows α"[1]. The state that agent 1 knows that agent 2 knows α is formalized by $K_1K_2\alpha$. A formula like $K_1\alpha \wedge K_1(\alpha \rightarrow \beta) \rightarrow K_1\beta$ is interpreted: "if agent 1 knows α and $\alpha \rightarrow \beta$ then he knows β".

Formally, the language $\mathcal{L}_N^K$ of modal epistemic logic is defined as follows:

Definition 1 (The language of epistemic logic) Let *Atom* be a nonempty, countable set of atomic formulae and $Agent = \{1, \ldots, N\}$ a set of agents. $\mathcal{L}_N^K$ is the least set such that

1. $Atom \subseteq \mathcal{L}_N^K$
2. If $\alpha \in \mathcal{L}_N^K$ then $\neg\alpha \in \mathcal{L}_N^K$
3. If $\alpha \in \mathcal{L}_N^K$ and $\beta \in \mathcal{L}_N^K$ then $(\alpha \rightarrow \beta) \in \mathcal{L}_N^K$
4. If $\alpha \in \mathcal{L}_N^K$ and $i \in Agt$ then $K_i\alpha \in \mathcal{L}_N^K$

The modal depth of a formula is defined by the following conditions: $depth(\phi) = 0$ for all $\phi \in Atom$; $depth(\neg\alpha) = depth(\alpha)$; $depth(\alpha \rightarrow \beta) = max(depth(\alpha), depth(\beta))$; and $depth(K_i\alpha) = depth(\alpha) + 1$.

2.1.2 Axioms for modal epistemic logic

A modal epistemic logic for N agents is obtained by joining together N modal logics, one for each agent. For simplicity's sake it is usually assumed that the agents are homogeneous, i.e., they can be described by the same logic. So an epistemic logic for N agents consists of N copies of a certain modal logic. Such a system is denoted by the same name as the modal system, but with the subscript N, e.g., $\mathbf{K}_N$ is the logic consisting of N copies of the logic **K**.

Definition 2 (Modal epistemic logic $\mathbf{K}_N$) $\mathbf{K}_N$ is the modal epistemic logic specified by the following axioms and rules of inference (where $i = 1, \ldots, N$):

(PC) All propositional tautologies

(K) $K_i\alpha \wedge K_i(\alpha \rightarrow \beta) \rightarrow K_i\beta$

(MP) Modus ponens: from α and $\alpha \rightarrow \beta$ to infer β

(NEC) From α to infer $K_i\alpha$

Stronger logics can be obtained by adding additional principles, which express the desirable properties of the concept of knowledge, to the basic system $\mathbf{K}_N$. The following properties are often considered:

[1] The truth values of epistemic statements also depend on other parameters such as time, location, context. However, it is a common practice in epistemic logic to take only agents into consideration and to assume certain standard values for the other parameters, i.e., the sentences are interpreted relative to the "current" situation. If only one agent is considered then even the reference to the agent is omitted.

(T) $K_i\alpha \to \alpha$

(D) $K_i\alpha \to \neg K_i\neg\alpha$

(4) $K_i\alpha \to K_iK_i\alpha$

(5) $\neg K_i\alpha \to K_i\neg K_i\alpha$

The formula **(T)** states that knowledge must be true. One normally takes this property to be the major one distinguishing knowledge from belief: you can have false beliefs, but you cannot know something that is not true. For that reason **(T)** is sometimes called the Knowledge Axiom or the Truth Axiom (for knowledge). Systems containing the schema **(T)** (such as $\mathbf{S4}_N$ and $\mathbf{S5}_N$) are then called logics of knowledge, and logics without the schema **(T)** are called logics of belief.[2]

The property **(D)**, occasionally called the Consistency Axiom, requires that agents be consistent in their knowledge: they do not know both a formula and its negation. Often the formula $\neg K_i(\alpha \wedge \neg\alpha)$ is used instead of **(D)**. These two formulae are equivalent in all logics containing $K_i\alpha \wedge K_i\beta \leftrightarrow K_i(\alpha \wedge \beta)$, in particular in all normal modal systems. Generally, **(D)** is a weaker condition than **(T)**.

The properties **(4)** and **(5)** are called positive and negative introspection axioms, respectively. They say that an agent is aware of what he knows and what he does not know. Their converses, i.e., the formulae $K_iK_i\alpha \to K_i\alpha$ and $K_i\neg K_i\alpha \to \neg K_i\alpha$, are instances of the schema **(T)**. Taking **(4)** and **(5)** together with their converses we have $K_iK_i\alpha \leftrightarrow K_i\alpha$ and $K_i\neg K_i\alpha \leftrightarrow \neg K_i\alpha$, which allow to reduce multiple knowledge operators to a single (positive or negative) knowledge operator.

The commonly used epistemic logics are specified as follows:

- $\mathbf{T}_N$ is $\mathbf{K}_N$ plus **(T)**
- $\mathbf{S4}_N$ is $\mathbf{T}_N$ plus **(4)**
- $\mathbf{S5}_N$ is $\mathbf{S4}_N$ plus **(5)**
- $\mathbf{KD}_N$ is $\mathbf{K}_N$ plus **(D)**
- $\mathbf{KD4}_N$ is $\mathbf{KD}_N$ plus **(4)**
- $\mathbf{KD45}_N$ is $\mathbf{KD4}_N$ plus **(5)**

That is, the logics $\mathbf{KD}_N$, $\mathbf{KD4}_N$ and $\mathbf{KD45}_N$ are obtained by substituting the axiom schema **(D)** for **(T)** in the axiomatization of $\mathbf{T}_N$, $\mathbf{S4}_N$ and $\mathbf{S5}_N$ respectively.

It is easily verified that the following inference rules are valid for $\mathbf{K}_N$ and its normal extensions:

(NEC) From α to infer $K_i\alpha$ (Necessitation)

(MON) From $\alpha \to \beta$ to infer $K_i\alpha \to K_i\beta$ (Monotony)

[2] It should be noted, however, that in AI terminology, no sharp distinction between knowledge and belief as in philosophy is made: knowledge is not required to be true. Unless stated otherwise I shall follow this terminology and use the term "knowledge" in the wider sense.

(CGR) From $\alpha \leftrightarrow \beta$ to infer $K_i\alpha \leftrightarrow K_i\beta$ (Congruence)

($\mathbf{RK}_n$) From $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ to infer $K_i\alpha_1 \wedge \ldots \wedge K_i\alpha_n \rightarrow K_i\beta$, for all $n \in \omega$

2.1.3 Possible-worlds semantics for epistemic logic

The intuitive idea behind the possible worlds approach is that an agent can build different models of the world using some suitable language. He usually does not know exactly which one of the models is the right model of the world. However, he does not consider all these models equally possible. Some world models are incompatible with his current information state, so he can exclude these incompatible models from the set of his possible world models. Only a subset of the set of all (logically) possible models are considered possible by the agent. For example, an agent possesses the information that he is 30 years old. Then among the models of the world he will not consider possible all those models in which he is not 30 years old. The smaller the set of worlds an agent considers possible, the smaller his uncertainty, and the more he knows.

The set of worlds considered possible by an agent i depends on the "actual world", or the agent's actual state of information. This dependency can be captured formally by introducing a binary relation, say R_i, on the set of possible worlds (read possible models of the world.) To express the idea that for agent i, the world t is compatible with her information state when he is in the world s, we require that the relation R_i holds between s and t. One says that t is an epistemic alternative to s (for agent i). If a sentence α is true in all worlds which agent i considers possible then we say that this agent knows α. Formally, the concept of models is defined as follows:

Definition 3 A model M for the language $\mathcal{L}_N^K$ comprises a nonempty set S of possible worlds (or states), N binary relations $R_1, \ldots, R_N$ on S (one for each agent), and a valuation function $V : Atom \mapsto Pow(S)$. The satisfaction relation $\models$ is defined recursively on $\mathcal{L}_N^K$ as follows:

- $M, s \models \phi$ iff $s \in V(\phi)$, for all atomic formulae $\phi \in Atom$
- $M, s \models \neg\alpha$ iff $M, s \not\models \alpha$, i.e., it is not the case that $M, s \models \alpha$
- $M, s \models \alpha \rightarrow \beta$ iff $M, s \not\models \alpha$ or $M, s \models \beta$
- $M, s \models K_i\alpha$ iff for all $t \in S$, sR_it implies $M, t \models \alpha$

The relations $R_1, \ldots, R_N$ are called relations of epistemic alternativeness, or accessibility relations. A formula α is said to be valid with respect to a class of models if for each model M in that class and each world $s \in S$ we have that $M, s \models \alpha$.

We can easily check that according to definition 3, if $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ is valid then so is $K_i\alpha_1 \wedge \ldots \wedge K_i\alpha_n \rightarrow K_i\beta$, for all $i \in Agent$ and all natural numbers $n = 0, 1, 2, \ldots$. These rules can be interpreted as saying that any agent i's knowledge is closed under logical laws: whenever i knows all premises of a valid inference rule then he also knows the conclusion.

If we restrict the class of models by imposing appropriate conditions on the epistemic alternativeness relations R_i's then we get larger classes of valid formulae and

may obtain characteristic models for extensions of $\mathbf{K}_N$. The well-known results for modal logic can be transferred to epistemic logic without any difficulty. The following theorem summarizes some completeness and decidability results for modal epistemic logic (cf. [Che80], [HC96], [Gol87], [HM92], [FHMV95]).

Theorem 4 1. $\mathbf{K}_N$ is determined by the class of all models with N accessibility relations.

2. $\mathbf{T}_N$ is determined by the class of models with N reflexive accessibility relations.

3. $\mathbf{S4}_N$ is determined by the class of models with N reflexive and transitive accessibility relations.

4. $\mathbf{S5}_N$ is sound and complete wrt the class of models with N equivalence relations as accessibility relations.

5. $\mathbf{KD}_N$ is determined by the class of models with N serial accessibility relations.

6. $\mathbf{KD4}_N$ is determined by the class of models where the N accessibility relations are serial and transitive.

7. $\mathbf{KD45}_N$ is sound and complete wrt the class of models where the N accessibility relations are serial, transitive, and Euclidean.

8. $\mathbf{K}_N$, $\mathbf{T}_N$, $\mathbf{S4}_N$, $\mathbf{S5}_N$, $\mathbf{KD}_N$, $\mathbf{KD4}_N$, and $\mathbf{KD45}_N$ are all decidable.

The logic $\mathbf{S5}_N$ is considered by many researchers as the standard logic of rational knowledge, and $\mathbf{KD45}_N$ as the standard belief logic. It is generally accepted that negative introspection is a more demanding condition than positive introspection. Therefore many researchers argue that it is more reasonable to adopt $\mathbf{S4}_N$, rather than $\mathbf{S5}_N$, as the logic of knowledge.

2.2 Adding common knowledge

Using the language $\mathcal{L}_N^K$ it is possible to express that some agent knows a certain fact, or an agent knows that another agent knows that he knows some fact, and so on. However, for a number of situations this language is not expressive enough: the state of knowledge in certain situations can only be described by an infinite number of iterations: everyone (in a group) knows simultaneously a fact α, everyone knows that everyone knows α, everyone knows that everyone knows that everyone knows α, and so on. In such a case we say that α is *common knowledge* among the group.

Common knowledge turns out to be a crucial concept in in explaining the rationality of certain actions, namely those co-operative enterprises such as conventional social practices, including language. It was first studied by David Lewis in the context of convention ([Lew69]), who observes that in order for something to be a convention, it must be common knowledge in the group. The notion has subsequently been applied to the analysis of language and discourse understanding ([Sch72], [CM81]), of games ([Aum76], [Gea92]), and of distributed systems ([Hal87]).

Current theories of intelligent agents usually take an agent-centric viewpoint, i.e., agents are viewed from the perspective of the designer of a single agent. Therefore, individual knowledge is of far greater interest than common knowledge. Nevertheless, the concept of common knowledge is of interest because it raises problems about the complexity of cognitive states which we can sensibly attribute to each other.

Although we could define common knowledge for each nonempty subset of the set *Agt* of agents, for simplicity we consider only common knowledge of the whole group. The language $\mathcal{L}_N^{CK}$ of epistemic logic with common knowledge is obtained by adding a new operator C to the language $\mathcal{L}_N^K$. The formula $C\alpha$ is interpreted as: "α is common knowledge of the agents". Formally, $\mathcal{L}_N^{CK}$ is defined as follows:

Definition 5 (The language of epistemic logic with common knowledge) Let *Atom* be a set of atomic formulae and *Agent* a set of agents as defined in Definition 1. $\mathcal{L}_N^{CK}$ is the least set such that

1. $Atom \subseteq \mathcal{L}_N^{CK}$
2. If $\alpha \in \mathcal{L}_N^{CK}$ then $\neg\alpha \in \mathcal{L}_N^{CK}$
3. If $\alpha \in \mathcal{L}_N^{CK}$ and $\beta \in \mathcal{L}_N^{CK}$ then $(\alpha \to \beta) \in \mathcal{L}_N^{CK}$
4. If $\alpha \in \mathcal{L}_N^{CK}$ and $i \in Agt$ then $K_i\alpha \in \mathcal{L}_N^{CK}$
5. If $\alpha \in \mathcal{L}_N^{CK}$ then $C\alpha \in \mathcal{L}_N^{CK}$

The auxiliary operator E (to be interpreted as "everyone knows") is defined as:

$$E\alpha =_{def} K_1\alpha \wedge \ldots \wedge K_N\alpha$$

Logics of common knowledge can be axiomatized on the basis of the corresponding epistemic logics by adding suitable axiom schemata and inference rules. The following axiomatization is due to Halpern and Moses ([HM92]).

Definition 6 (Systems of epistemic logic with common knowledge) Let Λ be one of the logics $\mathbf{K}_N$, $\mathbf{T}_N$, $\mathbf{S4}_N$, $\mathbf{S5}_N$, $\mathbf{KD}_N$, $\mathbf{KD4}_N$ and $\mathbf{KD45}_N$. Then Λ^C is the logic obtained by adding to Λ the following axiom postulates:

(FP) $C\alpha \to E(\alpha \wedge C\alpha)$ (Fixpoint axiom)

(RI) From $\alpha \to E(\alpha \wedge \beta)$ infer $\alpha \to C\beta$ (Rule of Induction)

Various other axiomatizations exist, e.g., by Kraus and Lehmann ([KL88]), Lismont ([Lis93]), Lismont and Mongin ([LM94]), and Bonanno ([Bon96]).

Logics of common knowledge can be given an adequate possible worlds semantics (cf., e.g., [KL88], [HM92], [FHMV95]). As in definition 3, each knowledge operator is interpreted by means of a binary relation on the set of possible worlds. An additional alternativeness relation is introduced to interpret the common knowledge operator. To capture the relationship of individual and common knowledge, it is stipulated that the relation corresponding to the common knowledge operator is the transitive closure of the union of the accessibility relations which correspond to the knowledge operators. Formally:

Definition 7 A model for the language $\mathcal{L}_N^{CK}$ with N agents is a structure $M = (S, R_1, \ldots, R_N, R^+, V)$ where S is a nonempty set, $R_1, \ldots, R_N, R^+$ are binary relations on S such that $R^+ = (R_1 \cup \ldots \cup R_N)^+$,and V is a valuation function $V : Atom \mapsto Pow(S)$. The satisfaction relation $\models$ is defined recursively as follows:

- $M, s \models \phi$ iff $s \in V(\phi)$, for all atomic formulae $\phi \in Atom$
- $M, s \models \neg\alpha$ iff $M, s \not\models \alpha$, i.e., it is not the case that $M, s \models \alpha$
- $M, s \models \alpha \rightarrow \beta$ iff $M, s \not\models \alpha$ or $M, s \models \beta$
- $M, s \models K_i\alpha$ iff for all $t \in S$, sR_it implies $M, t \models \alpha$
- $M, s \models C\alpha$ iff sR^+t implies $M, t \models \alpha$

A model is said to have a certain property if the accessibility relations $R_1, \ldots, R_N$ have that property. (Note that R^+ needs not necessarily have that property.) The following theorem ([FHMV95]) lists some well-known completeness results about logics of common knowledge.

Theorem 8

1. $\mathbf{K}_N^C$ is determined by the class of all models.
2. $\mathbf{T}_N^C$ is determined by the class of reflexive models.
3. $\mathbf{S4}_N^C$ is determined by the class of reflexive and transitive models.
4. $\mathbf{S5}_N^C$ is determined by the class of reflexive, transitive and symmetric models (i.e., $R_1, \ldots, R_N$ are equivalence relations.)
5. $\mathbf{KD}_N^C$ is determined by the class of serial models.
6. $\mathbf{KD4}_N^C$ is determined by the class of serial and transitive models.
7. $\mathbf{KD45}_N^C$ is determined by the class of serial, transitive, and Euclidean models.

2.3 Epistemic logic and agent theories

Relating an agent's beliefs and desires to its action has been one of the major challenges to practical reasoning, i.e., reasoning that we use to decide what to do. Practical reasoning has long been a field of philosophical studies. Examples include the study of various forms of the so-called practical syllogism:

$$\frac{\begin{array}{c}x \text{ wants } B \\ x \text{ knows that doing } A \text{ leads to } B\end{array}}{\text{Therefore } x \text{ does } A}$$

The logical analysis of practical reasoning was pioneered by G. H. von Wright ([vW63], [vW72]). Since the 1980s, AI research has seen a revival of interest in theories of knowledge and action. The relationship between knowledge and action are being investigated intensively in the field of intelligent agents research, and a number of

sophisticated theories have been proposed for describing this relationship. In this section I will examine briefly the role that the epistemic concepts play in some of the more influential agent theories. For an overview of recent agent theories consult [WJ95].

Modern theories of knowledge and action are built up from some basic mental, i.e., informational and motivational attitudes (like knowledge, belief, goal, intention), together with some "objective" modalities (like time, possibility, chance). By far, the latter concepts are much less controversial than the former ones. Formal theories of these "objective" modalities can be developed independently on any theory of mental concepts, but the converse is not necessarily true. For example, systems of modal or temporal logic do not presume any logic of mental notions, but a theory of intention is typically developed on the basis of some temporal logic.

As to the mental attitudes, there is no agreement about the choice of the set of the primitive notions. However, the informational aspect seems so fundamental that it is agreed that knowledge (in the sense of know-that) cannot be defined in terms of others and should be included as one of the basic concepts. On the other hand, the concept of knowledge is essential in theories of other mental notions like intention, know-how, or even desire and goal. For example, know-how is normally defined in terms of knowledge: knowing how to achieve a goal includes the knowledge that after doing something, certain facts will obtain.

The first formalizations of knowledge and action in AI was carried out in the late 1970s and early 1980s. The primary interest was to study knowledge as pre-condition for executing plans. Inspired by ideas of McCarthy and Hayes ([MH69]), Robert Moore developed a formal theory of knowledge which is essentially modal logic **S4**, but expressed in the first-order metatheory ([Moo90]).

More recently, Cohen and Levesque's theory of intention ([CL90]) has been very influential. Following Bratman's analysis of intention and the role that intentions play in human practical reasoning ([Bra87], [BIP88]), Cohen and Levesque identify the key properties that must be satisfied by a reasonable theory of intention. They develop a formal theory based on two primitive mental notions: belief and goal. The logic of belief is assumed to be the modal system **KD45**, and that of goal **KD**. Together with two (temporal) modalities indicating that some event will happen next and some event has just happened, they are able to define the concept of intention and to show that many of Bratman's requirements for a theory of intention are satisfied.

In another attempt to formalize Bratman's theory of intention, Rao and Georgeff ([RG91b], [RG91a]) have developed a logical framework for agent theory based on three primitives: belief, desire, and intention. Within this BDI (Belief - Desire - Intention) architecture, belief is treated as a basic modality which satisfies the **KD45** axioms. Desire and intention are assumed to be **KD**-modalities. The BDI architecture has been adopted and further developed subsequently by a number of researchers ([GR95], [Sin94], [Sin95], [Woo96]).

In related work to formalize properties of intelligent agents, Meyer et. al. have proposed the KARO (Knowledge - Abilities - Results - Opportunities) architecture ([vdHvLM94], [vLvdHM94]). In this architecture, **KD45** is assumed as the logic of belief, and **S5** is used to formalize knowledge.

Although not strictly a logic-based theory of agency, the AOP (Agent oriented programming) paradigm ([Sho93]) also deals with the behavior of rational agents. Again, belief is taken as one basic mental concept and is formalized using the modal logic **KD45**. Moreover, belief is also used to characterize commitment (or obligation), another basic mental concept: besides the **KD**-axioms, the concept of commitment must also satisfy some additional rationality postulates, which basically say that commitments are known.

To summarize, the most influential among the recent agent theories are developed on the basis of modal epistemic logic. Now I shall argue that the modal approach is not suitable because it does not yield specifications of cognitive states which can play a justificatory role for agents' action. The agent model provided by modal epistemic logic does not accord with generally agreed facts about the nature of intelligent agents, in particular with the fact that they are limited in the amount and complexity of the information they can handle.

2.4 The problem of logical omniscience

The treatment of epistemic logic as a branch of modal logic brings some advantages. However, there is a high price to pay. The most important objection to the modal approach is that it makes unrealistic assumptions about the reasoning power of the agents. The problem is known as the "logical omniscience problem" (LOP) and occurs in several forms. In its strongest form the problem can be stated as follows:

Lemma 9 Let Λ be any normal modal logic containing $\mathbf{K}_N$. For any $\Gamma \subseteq \mathcal{L}_N^K$, $\alpha \in \mathcal{L}_N^K$, and $i \in \mathit{Agent}$, if $\Gamma \vdash_\Lambda \alpha$ then $K_i(\Gamma) \vdash_\Lambda K_i\alpha$, where $K_i(\Gamma) =_{def} \{K_i\gamma : \gamma \in \Gamma\}$.

That is, whenever an agent knows all of the formulae in a set Γ and α follows logically from Γ, then the agent also knows α. In particular, the agent knows all theorems (taking Γ in lemma 9 to be the empty set), and he knows all logical consequences of a sentence that he knows (taking Γ to consist of a single sentence.)

Besides this strong form there are other, generally weaker forms of logical omniscience. The following are listed in [FHMV95]:

- Knowledge of valid formulae: agent i knows all logical truths (rule **(NEC)**).
- Closure under logical implication: if agent i knows α and if α logically implies β (i.e., $\alpha \rightarrow \beta$ is valid), then agent i knows β (rule **(MON)**).
- Closure under logical equivalence: if agent i knows α and if α and β are logically equivalent (i.e., $\alpha \leftrightarrow \beta$ is valid), then agent i knows β (rule **(CGR)**).
- Closure under material implication: if agent i knows α and if agent i knows $\alpha \rightarrow \beta$ then agent i knows β (axiom **(K)**).
- Closure under conjunction: if agent i knows α and if agent i knows β then agent i knows $\alpha \wedge \beta$ (axiom **(C)**).

The list of questionable properties could be extended to include any other instance of the rule **(RK$_n$)** (from $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ to infer $K_i\alpha_1 \wedge \ldots \wedge K_i\alpha_n \rightarrow K_i\beta$.) Moreover, the axiom schemata **(D)**, **(4)** and **(5)** can also be shown to be too strong for realistic agents. In particular, under certain circumstances axiom **(5)** suggests that agents can even decide undecidable problems ([BS92], [SW94])! In general, there seems to be no genuine epistemic principle that may claim universal validity[3].

If epistemic logic is to be interpreted as describing actual knowledge of realistic (though idealized) agents, then the discussed closure properties require agents to be very powerful reasoners whose computational capacities cannot be achieved by real (human or artificial) agents, who are simply not logically omniscient. Logical omniscience poses a problem because it contradicts the fact that agents are limited in their reasoning powers. They are inherently resource-bounded and therefore cannot handle an unlimited amount of information. Agents may establish immediately certain logical truths or simple consequences of what they consciously assented to. However, there are highly remote dispositional states which could only be established by complex, time-consuming reasoning. The modal framework cannot distinguish between a sentence that an agent consciously assented to and a piece of potential knowledge which could never be made actual by the agent and is therefore not suited to model resource-bounded reasoning[4].

2.4.1 Implicit knowledge

The LOP shows that the view that modal epistemic logic describes actual knowledge of idealized agents is not tenable. A certain degree of idealization is meaningful and even necessary in philosophical and scientific research. However, the idealizations made by modal epistemic logic are so strong that the agents they describe have hardly anything in common with real agents. Those "agents" are merely theoretical constructs without any empirical basis. Hence, the modal approach is not suited to capturing the notion of actual knowledge (or belief) adequately.

But if modal epistemic logics do *not* describe what agents actually know, what *do* they describe then? Well, they can be interpreted as logics of a related, but different concept. It is remarked by several authors that the laws of modal systems are acceptable if the formula $K_i\alpha$ is read "agent i knows α implicitly" ([Lev84], [FH88],) "α follows from i's knowledge" ([FHMV95]), "agent i carries the information α" ([Bar89]), or "α is agent i's possible knowledge" ([HK91]), instead of "the agent i knows α". Although the technical definitions may differ, "implicit knowledge" and similar terms are all used in the same spirit: they refer to what is implicitly represented in an agent's

[3]Some formulae containing the knowledge operators are always valid, but they are not genuine epistemic statements: a formula like $K_i\alpha \rightarrow K_i\alpha$ does not say anything about an agent's reasoning capacities. Counterexamples to the commonly assumed epistemic closure principles are presented in [Len78], [Hal95], among others.

[4]If the possible-worlds semantics is adapted for modeling motivational concepts such as goals, desires, or intentions, then the resulting logics suffer from a similar problem: if an agent intends to do something then he intends all logical consequences of his intention. This is not a desirable property: one might intends to go to the dentist without having an intention of suffering pain, although the latter is a necessary consequence of the former. This problem is known as the side-effect problem (cf. [Bra90]).

information state, i.e., what logically follows from his actual knowledge. They describe dispositional states which could only be established by reasoning and reflection upon one's mental states. The concept of implicit knowledge is used without any notion of agents computing knowledge or having to answer questions based on their knowledge. What an agent actually knows is called his explicit knowledge.

If "knowledge" is understood as "implicit knowledge", then the forms of logical omniscience discussed previously are no longer a problem: although the discussed axioms and inference rules are not reasonable for the explicit view, they are acceptable for the implicit view of knowledge. Thus, modal epistemic logics seem to be acceptable for the purpose of formalizing the concept of implicit knowledge. They should be interpreted as logics of implicit, or potential knowledge, and not as logics of explicit, or actual knowledge.

From the viewpoint of agent theories, actual (explicit) knowledge is clearly more important than implicit knowledge: it is the former kind of knowledge that agents can act upon, but not the latter. The mere implicit knowledge that some path connecting all towns in a region is the shortest one is useless for a traveling salesman who seeks to maximize his profit — he must make this implicit knowledge explicit in order to choose what path to travel. The implicit, but not explicit knowledge of a winning strategy is useless for a chess player who must make the next move within a short time. An information agent whose knowledge is represented as a knowledge base must normally make complex and time-consuming inferences before he can answer a query.

Since agents need to act on the basis what they *actually* know, and not what they merely *potentially* know, agent theories must be based on logics that can capture what agents actually know. Because of the importance of explicit knowledge for agents' action, the search for logics of explicit knowledge has been continuing, and a number of systems have been proposed for that purpose. In the next chapter I shall review the most important attempts to model explicit knowledge and show why they are not suitable as a basis for agent theories.

Other models of knowledge

Any adequate theory of agency must be able to describe explicit knowledge correctly, because this is the concept of knowledge that can provide justification and explanation for action. What an agent chooses to do depends on his explicit, and not on his implicit knowledge. As I have argued in the last chapter, any theory of explicit knowledge must avoid logical omniscience. Because modal epistemic logic cannot characterize explicit knowledge, a number of alternative approaches have been proposed.

In this chapter I shall examine some of the more prominent attempts to develop logics for explicit knowledge of non-omniscient agents. I shall show that although the existing approaches to explicit knowledge can solve the logical omniscience problem, they are not suited to characterizing the information states of agents. In order to be useful, a logic of explicit knowledge must satisfy some additional conditions besides the lack of omniscience. This explains why all influential agent theories use modal epistemic logics for formalizing the informational aspect of agents, although the modal approach does not capture the explicit notion of knowledge adequately.

3.1 Logics for non-omniscient agents

The lack of logical omniscience can stem from various sources. An agent may not be aware of a sentence and therefore does not know it. He may be restricted in his logical capabilities and does not know all the axioms and inference rules. Or he may be biased and refuses to use certain rules of inference. It is also possible that an agent does not care about the consequences of a sentence, so he does not even try to compute them. However, the most important source of non-omniscience is the agents' resource boundedness: they simply do not have enough computational capacities (time, memory etc.) to compute all the consequences of their knowledge, even if all inference rules are available. It is not difficult to supply an agent with a sound and complete deduction mechanism, especially in the context of artificial intelligent agents. Such agents are not omniscient simply because they are resource bounded.

By exploiting the different sources of non-omniscience it can be possible to model non-omniscient agents. For example, by demanding that knowledge include awareness one can describe agents who are not logically omniscient because they are not aware of some formulae. By restricting the set of admissible inference rules that can be used by agents one can model agents who are not able to or refuse to use certain inference rules. Another way to develop a model of knowledge and belief based on the

resource-bounded inferential capabilities of agents is to stipulate that the agents can only compute formulae whose derivations require at most n inference steps, for some fixed value of n.

3.1.1 Weak deduction mechanisms

The obvious strategy to solve the logical omniscience problem is to weaken epistemic logic. One denies the universal validity of the mentioned inference rules **(NEC)**, **(MON)**, and **(CGR)**, or one of the essential axioms like **(K)**. In fact, almost all attempts to solve the LOP have in common that they consider (families of) systems that are weaker than the standard modal epistemic logics in the following sense. Firstly, not all theorems of modal epistemic logic are provable in those systems. Secondly, the set of formulae known by an agent at a state is not necessarily closed under the laws of the propositional calculus: a formula β may be provable from $\alpha_1, \ldots, \alpha_n$ by using axioms and inference rules of the propositional calculus, but $K_i\beta$ cannot be derived from $K_i\alpha_1, \ldots, K_i\alpha_n$ within the epistemic logic under consideration.

To construct such weak systems we can postulate, for example, that the agent only knows some "obvious" logical truths, but not necessarily the "more complicated" ones. We can assume that the agent can draw all "obvious" consequences, but not any arbitrary consequence of a certain sentence. This is achieved by postulating that the deduction mechanism of the agents is not complete, that is, it is not powerful enough to allow the agents to draw all logical consequences of their knowledge ([Hin70], [Ebe74], [Ste84], [Kon86], [Wut91], [GG93]). If an agent's inference mechanism is kept very weak, then logical omniscience could be avoided.

Certain systems of modal logic are able to characterize agents whose inference mechanisms are weaker than propositional consequence. For example, non-normal systems can be used for describing an agent who does not know all logical truths, and non-monotonic modal logics can model agents whose knowledge is not closed under logical consequence[1]. If a very weak modal logic (e.g., a classical system) is employed to model knowledge, then most versions of the LOP are solved: neither the necessitation rule **(NEC)**, nor the monotony rule **(MON)**, nor the axiom schema **(K)** is valid in a weak classical system. However, some weaker versions of the LOP still remain unsolved. All classical modal logics are closed under the congruence rule **(CGR)**, so an agent described by such a modal system knows all logical equivalences of a sentence that he knows. Such a closure property is obviously too strong for real agents.

Another group of attempts to gain control over the LOP is to consider nonstandard logics to model agents' reasoning. The intuitive idea is as follows. Modal epistemic logic assumes that agents use classical logic (or more accurately, some extension of classical logic) in their reasoning. This causes logical omniscience because the notion of logical consequence defined by classical logic is too powerful, i.e., too much can be inferred from some base of knowledge. In particular, all tautologies are known

[1] Normal modal logics are systems which are closed under the (knowledge) necessitation rule **(NEC)**, and monotonic modal logics are closed under the monotony rule **(MON)**. In the context of modal logic, the term "monotonic" means that the rule **(MON)** holds. This usage should not be confused with the terminology of non-monotonic reasoning research. Consult appendix A for a brief overview of modal systems weaker than the minimal normal modal logic **K**.

because classical logic allows to derive them from the empty set. Hence, if the notion of logical consequence is restricted so that not all classical consequences can be drawn then certain forms of omniscience can be avoided. Such a restriction can be achieved by employing a nonstandard logic. Among the non-classical logics that have been employed for that purpose are several variants of relevance logic ([Lev84], [FHV95]) and many-valued logics ([Ho93]).

Although the approaches based on nonstandard logics solve certain forms of the LOP, they cannot eliminate the LOP completely. The agents described by those logics are not logically omniscient wrt classical logic, but they are omniscient wrt to some (nontrivial) non-classical logic. Such attempts cannot be considered satisfactory solutions to the LOP. Consequently, they are not suitable for characterizing explicit knowledge. In general, any logic that cannot model what is explicitly available to the agents but only information that must be inferred using some — possibly incomplete — deduction mechanism must be viewed as a logic of implicit knowledge[2].

3.1.2 Impossible possible worlds

A number of systems have been proposed which assume still more restricted reasoning capacities of the agents and in this way avoid all forms of logical omniscience. One framework that eliminates logical omniscience completely is the so-called impossible-worlds approach. Logical omniscience can be avoided if one allow "impossible possible worlds" in which the valuation of the sentences of the language is arbitrary. In other words, the logical laws do not hold in the "impossible possible worlds" ([Cre70], [Cre73], [Hin75], [Ste79], [Ran82], [Wan90]).

The intuition underlying the introduction of impossible worlds is that an agent may regard some models of the (real) world possible, although they are logically impossible. For example, a logical contradiction cannot be true. However, an agent may not have enough resources to determine the truth value of that contradiction and simply assumes it to be true. So he will consider some worlds possible, although logically they are impossible.

Definition 10 (Impossible-worlds structures) An impossible-worlds model for the language $\mathcal{L}_N^K$ is a tuple $M = (S, W, R_1, \ldots, R_N, V)$ where S is a nonempty set (the set of worlds), $W \subseteq S$ is the set of possible worlds (elements of $S \setminus W$ are called impossible worlds), $R_1, \ldots, R_N$ are binary relations on S, and $V : \mathcal{L}_N^K \times S \mapsto \{1, 0\}$ is a function that assigns arbitrary truth values to formulae of the language $\mathcal{L}_N^K$ in impossible worlds which behaves standardly on possible worlds, i.e., if $s \in W$ then:

- $V(\neg\alpha, s) = 1$ iff $V(\alpha, s) = 0$
- $V(\alpha \to \beta, s) = 1$ iff $V(\alpha, s) = 0$ or $V(\beta, s) = 1$
- $V(K_i\alpha, s) = 1$ iff $V(\alpha, t) = 1$ for all $t \in S$ such that sR_it.

[2]Strictly speaking, Levesque's logic of "explicit" belief ([Lev84]) still describes a kind of implicit belief, because what is defined to be explicit belief of an agent in that model is *not* immediately available to the agent. The same criticism applies to other models which intend to model explicit belief but still fall prey to some form of logical omniscience, e.g., Konolige's deduction model [Kon86], or the notion of algorithmic knowledge of Halpern et. al. ([HMV94]).

Satisfaction is defined as: $M, s \models \alpha$ iff $V(\alpha, s) = 1$. Validity is defined with respect to possible worlds only: α is valid wrt impossible-worlds models iff for each impossible-worlds model M and possible world $s \in W$ we have $M, s \models \alpha$.

Because knowledge is evaluated with respect to all states and the laws of logic do not hold in some states, all forms of logical omniscience are avoided. For instance, the tautology $\alpha \vee \neg\alpha$ may be false in an impossible world, but an agent may consider that world possible, so $K_i(\alpha \vee \neg\alpha)$ does not hold universally. In other words, the necessitation rule is not valid. Similarly, axiom **(K)** (closure under material implication) fails to hold, because it is possible that in an impossible world both formulae α and $\alpha \to \beta$ are true while β is false.

The logic determined by the class of all impossible-worlds models is rather uninteresting, because no genuine epistemic statement is universally valid. Epistemic principles can be obtained by imposing appropriate conditions on the models. For example, axiom **(K)** is valid if for every impossible world, if the value 1 is assigned to both α and $\alpha \to \beta$ then it must be assigned to the formula β as well.

3.1.3 Awareness

Another solution to the LOP consists in introducing a new operator of awareness into the language and to require that belief include awareness ([FH88].) The underlying intuition is that agents need to be aware of some concept before they can have beliefs about it: one cannot know something one is completely unaware of. On the other hand, if an agent is aware of a formula α and implicitly knows α, then he knows α explicitly. The notion of awareness is left unspecified. Some possible interpretations of "agent i is aware of α" are: "i is familiar with all the propositions mentioned in α", "i is able to figure out the truth of α", or "i is able to compute the truth of α within time T."

For better comparison with other approaches, my presentation of the awareness framework will not follow the original definition ([FH88]) in details. The main intuitions are retained, however. In particular, there are no modal operators for implicit knowledge and awareness. The knowledge operators of the language $\mathcal{L}_N^K$ are now interpreted as explicit knowledge and will be evaluated accordingly in the definition of models.

Definition 11 (Awareness structures) An awareness model for the language $\mathcal{L}_N^K$ is a tuple $M = (S, R_1, \ldots, R_N, \mathcal{A}_1, \ldots, \mathcal{A}_N, V)$ where $M = (S, R_1, \ldots, R_N, V)$ is a Kripke model in the sense of definition 3, and $\mathcal{A}_i : S \mapsto Pow(\mathcal{L}_N^K)$ is a function associating a set of formulae with each state, where $i = 1, \ldots, N$. The semantics for atomic formulae, negations, and implications is as usual (cf. definition 3.) The clause for formulae of the form $K_i\alpha$ becomes:

- $M, s \models K_i\alpha$ iff $\alpha \in \mathcal{A}_i(s)$ and for all $t \in S$, if sR_it then $M, t \models \alpha$

Intuitively, $\mathcal{A}_i(s)$ is the set of formulae that agent i is aware of at state s, and the relations $R_1, \ldots, R_N$ are used to model implicit knowledge. The set of formulae that an agent is aware of can be arbitrary and needs not be closed under any law. Moreover, there is no relationship between (implicit) knowledge and awareness at all: the function

$\mathcal{A}_i$ and the relation R_i are completely independent. Since explicit knowledge is defined as implicit knowledge plus awareness, it is obvious that if an agent is aware of all formulae of the language then explicit knowledge reduces to implicit knowledge.

Because it is possible that an agent is aware of some sentence but he is not aware of its logical consequences or its equivalent sentences, the theorems and inference rules of modal epistemic systems do not hold in general. So the forms of logical omniscience discussed in chapter 2 are avoided.

That the awareness approach is able to model non-omniscient agents can be seen in another way. We have seen earlier that the impossible-worlds approach avoids all forms of logical omniscience. The following theorem shows that although the intuitions are quite different, the impossible-worlds approach and the awareness approach are equivalent in a precise sense (cf. [Wan90], [Thi93], [FHMV95]).

Theorem 12

- Let $M = (S, W, R_1, \ldots, R_N, V)$ be an impossible-worlds model. Then there is an awareness model $M' = (W, R'_1, \ldots, R'_N, \mathcal{A}_1, \ldots, \mathcal{A}_N, V')$ such that for any formula α of the language $\mathcal{L}_N^K$ and any $s \in W$ we have that $M, s \models \alpha$ iff $M', s \models \alpha$
- Let $M = (S, R_1, \ldots, R_N, \mathcal{A}_1, \ldots, \mathcal{A}_N, V)$ be an awareness model. Then there exists an impossible-worlds model $M' = (S', S, R'_1, \ldots, R'_N, V')$ such that for any formula α of the language $\mathcal{L}_N^K$ and any $s \in S$ we have that $M, s \models \alpha$ iff $M', s \models \alpha$

As an immediate consequence of this theorem, the awareness framework also solves all forms of the LOP: if an undesirable property can be falsified in an impossible-worlds model, then it can also be falsified in an awareness model. In fact, it can be seen easily that the set of $\mathcal{L}_N^K$-formulae which are valid wrt all awareness models consists of exactly the instances of propositional tautologies. In other words, no genuine epistemic statement is valid with respect to the class of all awareness models.

So far the concept of awareness has been left unspecified, so no meaningful restrictions can be placed on the set of formulae that an agent is aware of. Once a concrete interpretation has been fixed, some closure properties can be added to the awareness function to capture certain types of "awareness".

For example, if we consider a computer program that never computes the truth of a formula unless it has computed the truth of all its subformulae, then we may assume that awareness is closed under subformulae, i.e., if $\alpha \in \mathcal{A}_i(s)$ and β is a subformula of α then $\beta \in \mathcal{A}_i(s)$. This assumption may seem innocuous at first, but it turns out to have a rather strong impact on the properties of explicit knowledge. It can be shown easily that if awareness is closed under subformulae then an agent's knowledge is closed under material implication, i.e., the schema **(K)** is valid. In general, whenever β follows logically from $\alpha_1, \ldots, \alpha_n$ and β is a subformula of one of $\alpha_1, \ldots, \alpha_n$, then $K_i\beta$ follows from $K_i\alpha_1, \ldots, K_i\alpha_n$, for any agent i.

Another possible closure property for awareness is that agent might be aware of only a subset X of the atomic formulae. In this case one could assume that $\mathcal{A}_i(s)$ consists of exactly those formulae that are built up from the atomic formulae in X. Under this assumption some forms of logical omniscience are avoided, e.g., knowledge of valid formulae or closure under logical implication. However, all forms of the LOP occur again when attention is restricted to the sublanguage generated by X.

3.2 Logical omniscience vs. logical ignorance

In the previous section several attempts to develop logics of explicit knowledge have been reviewed. The proposed approaches try to avoid logical omniscience by considering agents with less deductive powers than those suggested by modal systems. By weakening epistemic logic the LOP can be solved, at least to some extent, and an agent's explicit knowledge can be described more realistically. Weak epistemic logics can be used to describe agents with very restricted reasoning capacities. In fact, many of the discussed approaches can even model agents who know nothing ("total ignoramusses"), those who cannot draw even the most elementary inferences ("complete idiots"), and those who believe nothing but contradictions ("ultimate fools").

Those irrational agents are clearly not very interesting. To describe more intelligent agents, the common way is to postulate axioms which describe the regularities of an agent's knowledge. Such axioms usually require that an agent's belief set, i.e., the set of formulae that he believes, is closed under certain logical laws. In this way the intuitive idea that the agent under consideration is somehow rational could be captured. The epistemic axioms are generally of the form: if all premises of a certain valid inference rule are known, then the conclusion is known. (This is also the general form of a theorem of a modal epistemic logic.) The more axioms are postulated, the more rational is the agent. In this way subsystems of the logic $\mathbf{K}_N$ can be obtained which do not suffer from the LOP and may describe agents more realistically than the modal systems. So, existing logics of explicit knowledge typically contain a subset of the axioms and rules for knowledge in the system $\mathbf{K}_N$, while other rules are rejected.

The strategy of employing weak epistemic logics for describing explicit knowledge can solve the logical omniscience problem, at least to some extent. However, other serious problems arise. Here I shall not discuss in details the specific problems of the various frameworks or try to solve them. I shall rather present a more fundamental criticism of the strategy of weakening epistemic logic and discuss the problems which arise when this strategy is pursued.

The use of subsystems of normal modal logics to describe explicit knowledge depends on the following assumption. Although it normally takes some effort to make a piece of implicit knowledge explicit, there are "obvious" logical consequences that should be recognized easily by rational agents. Therefore, it may be supposed that an agent's knowledge set at a time is always closed under those principles, although it is not closed under all logical laws. In other words, only closure properties corresponding to the "simple" inferences are regarded valid. So, the task of developing logics of explicit knowledge involves that of identifying "obvious" tautologies and logical inferences.

This assumption is far from plausible. However weak the epistemic postulates may be, they may still be too strong, at least for some agents. Moreover, a single axiom may seem innocuous, but joining it with other axioms may result in a rather powerful deductive system — and if a sentence can only be deduced by means of a powerful logic, then it is hardly justifiable to call it explicit knowledge. But even if the above assumption is accepted, many problems remain to be solved.

The first challenge is to select a set of postulates for knowledge which may be assumed to be valid for any rational agent. This is not at all an obvious choice. There is no objective, generally accepted criterion for deciding which tautologies are obvious,

which inferences are simple. Many criteria for identifying obvious consequences could be considered (and have been proposed). For example, one might maintain that obvious tautologies should be provable in less than n steps, where n is a fairly small natural number. One could restrict the modal depth of knowledge formulae to a small number. One could also demand that only consequences that are built up from subformulae of the premises can be drawn. However, none of these criteria is wholly convincing. The proof length is not a suitable measure for the simplicity of a tautology because it depends on the syntactical system being used. Many theorems have modal depth 1 and satisfy the subformulae condition, but they are still far from obvious, so neither the modal depth nor the subformulae condition is an adequate criterion. Therefore it is not possible to draw a sharp line between "simple rules" that should be usable by all rational agents and "complicated inferences" which cannot be assumed to be valid.

Another challenge is to find an appropriate way for modeling non-omniscient agents without making them logically ignorant. In order to account for the resource-boundedness of agents, their reasoning powers must be kept reasonably weak. However, if the use of certain inference rules is denied then the resulting logics may become too weak for many applications. Surely, logical omniscience must be avoided. But at the same time we are interested in having epistemic logics which are strong enough to allow sufficiently many conclusions to be drawn from a given set of facts about an agent's propositional attitudes. To interact with other agents, an agent needs to make assumptions about their rationality, and he should be able to assume that they are not logically ignorant. We want to model agents who know at least a large class of logical truths, and can draw sufficiently many conclusions from their knowledge.

The dilemma between logical omniscience and logical ignorance explains why modal epistemic logics are still widely used in agent theories despite the facts that implicit knowledge is useless when agents need to act *and* logics of explicit knowledge are readily available. The existing logics of explicit knowledge are not suited to characterizing agents because we want to model *rational, intelligent* agents, and not "complete idiots". They avoid logical omniscience, but they cannot offer anything what can account for the rationality of agents. Surely agents are not perfectly rational, yet they *are* rational. Facing the choice between "perfectly rational agents" and "complete idiots", agent theorists understandably opt for the former and use logics of implicit knowledge for modeling their agents, hoping that such logics can describe "almost correctly" what agents actually know.

The assumptions underlying the use of modal epistemic logics may be justified in some simple domains ("small worlds", "toy examples"), where the reasoning tasks involved are quite simple, where the decision process is not very complex, or when the time available is unlimited. In such simple domains, it can be assumed that whenever an agent needs some (implicitly available) information, he can perform the necessary inferences to have the information explicitly. However, such an assumption is not justified in more complex applications. Agents normally have to act under tight time constraints, their decisions what actions to be performed depend strongly on their actual knowledge, and the reasoning needed for making correct choices can be very complex and time-consuming. For example, calculating the shortest tour linking all towns in a region, computing the winning strategy in chess, and inferring the answer to a query from a given database are all very hard problems. It is obvious that modal

epistemic logics and other logics of implicit knowledge cannot describe correctly what agents actually know in such applications. To describe agents realistically in knowledge-intensive applications, we simply need other logics of knowledge.

What properties should a logic of knowledge have if it is to be useful for describing realistic, implementable agents? The first obvious requirement is that it must not suffer from the LOP. That is, it must not make unrealistic assumptions about the computational capabilities of agents. An epistemic principle can be regarded to be realistic if it can be confirmed empirically. For our purposes we shall employ the criterion that an agent can be implemented which constitute a model of the principle. Because agents can only handle a limited amount of information, we shall deal with agents whose explicit knowledge can be represented as a finite set of formulae and whose reasoning mechanism contains a finite number of inference procedures. This finiteness condition ensures that agents can be implemented.

Solving the LOP is necessary, but not sufficient for making a logic suitable for reasoning about knowledge. There are other requirements that must be fulfilled. It is important that the logic can do justice to the intuition that agents are rational: although the agents do not automatically know all consequences of their knowledge, they are in principle able to do so. Because of this rationality the agents are able to act upon their knowledge: they can answer questions based on their knowledge, they can plan their actions in advance, they can predict what other agents can and will do, and so on. If a logic cannot account for the agents' rationality, then there is hardly any justification at all to call it a logic of knowledge.

Another important requirement is that the logic be expressive enough to formalize "interesting" situations. This condition must remain somewhat vague, because different applications will require different expressive powers of the logic. However, we should keep in mind that the complexity of a logic generally increases with its expressive power, so we must try to find a good trade-off between expressiveness and simplicity.

The next two chapters describe some ways to model agents which are neither logically omniscient nor logically ignorant. In chapter 4 I shall show how explicit knowledge can be modeled without restricting the agents' rationality arbitrarily by denying them the use of certain inference rules. For modeling resource-bounded reasoning, what should be restricted is not the number of admissible inference rules, but the number of times they can be applied. I will show in chapter 5 how epistemic logic can be combined with a complexity analysis to describe resource-bounded reasoning more accurately.

Explicit knowledge

In the last chapter I have reviewed some prominent attempts to model the notion of explicit knowledge and discussed their main problems. In my opinion, the existing approaches fail to capture explicit knowledge adequately because they try to model entailment relations where none exists, namely within the set of sentences known by an agent at a single time point. Those attempts are doomed to failure because an agent's explicit knowledge at a time is simply not closed under logical laws and therefore cannot be described by any nontrivial logic. Forcing regularities upon an agent's explicit knowledge to make reasoning about it possible is not the proper way to cope with the difficulties.

In the following I shall suggest a new approach to reasoning about explicit knowledge which overcomes the drawbacks of existing approaches. The idea is to consider the evolution of one's knowledge over time: at one moment an agent may or may not know (explicitly) a certain consequence of his knowledge; however, he can perform some reasoning steps to know it at some moment in the future. I have argued that the traditional approaches fail to capture the concept of actual knowledge correctly because they do not take the cost of inferring new information into account: they assume that whenever an agent knows all premises of a valid inference rule then he automatically knows the conclusion. I will argue that axioms for epistemic logics must have the form: "if the agent knows all premises of a valid inference rule, and if he performs the correct inference step, then he will know the conclusion". In section 4.1 I shall discuss the main intuitions of my approach. Then, in section 4.2 formal systems will be defined and discussed.

4.1 The dynamics of knowledge

4.1.1 Explicit knowledge and reasoning actions

Let us consider an inference rule, say R. It can be a valid inference rule of classical logic, or some other (non-classical) logic, for example, intuitionistic logic, conditional logic or relevant logic. Assume that an agent accepts R as valid and he can use R. What does it mean? In the modal approach we formalize this idea by an axiom saying that the knowledge set of the agent is closed under this rule, that is, if all premises of the rule are known then the conclusion of R is also known. However, as we noted before, it is only true of implicit knowledge. In the context of explicit knowledge it

must mean something different. It means rather that, if the agent knows all premises of the rule, and if he perform the inference according to the rule R, then he will know the conclusion. The agent does not know the conclusion automatically, but rather as the result of some action, viz. the (mental) action of performing the corresponding inference. If he does not perform this action, then we cannot require him to know the conclusion, although this conclusion may seem to be an obvious consequences of the sentences under consideration.

The same line of argumentation applies to logical axioms, which can be viewed as inference rules without any premises. We cannot require an agent to know all axioms automatically and permanently, he must rather carry out some action before he can acquire knowledge of a certain axiom. Gaining knowledge of other, less obvious theorems is even harder: agents usually need to perform more complex computations in order to establish a theorem. Thus, it is possible that the agent knows all logical truths, but merely in principle. This knowledge is only implicit. In reality he never knows them all at once explicitly.

For formalizing the reasoning actions it is natural to use (a form of) dynamic logic ([Har84], [Gol87], [KT90]; see also appendix A for a brief overview.) We can add a set of basic actions to the language of epistemic logic. The set of formulae now includes formulae like $[R_i]K_i\alpha$ or $\langle R_i\rangle K_i\alpha$ with the intended meaning: "always after using rule R (or sometimes after using R) the agent i knows α". The formalization of the idea that an agent accepts and is able to use an inference rule is straightforward. For example, the idea that the agent i accepts modus ponens can be formalized by the axiom: $K_i\alpha \wedge K_i(\alpha \to \beta) \to \langle MP_i\rangle K_i\beta$. This axiom says no more than if agent i knows α and he also knows that α implies β, then after a suitable inference step he will know β.[1]

As the axioms can be viewed as special inference rules we can introduce an action corresponding to each agent and each axiom of the basis logic, which describes the ability of the agent to use this axiom in his reasoning. (In general, different agents may have different logics, so that the sets of basic actions are different for different agents. However, we assume a set of homogeneous agents, for the sake of simplicity.) By means of the familiar program connectives for dynamic logic (such as composition or iteration) we can formalize the idea that the agent may know the consequences of some sentence which he already knows explicitly, provided that he performs the right reasoning steps. For example, assume that the agent i knows the conjunction of α and $\alpha \to \beta$, that is, $K_i(\alpha \wedge (\alpha \to \beta))$. In all normal modal systems we can deduce $K_i(\alpha \wedge \beta)$. However, this inference is not sound for actual knowledge of realistic agents. There is no guarantee that the agent will know $\alpha \wedge \beta$ automatically, as the modal approach suggests. We can only say that *if* the agent reasons correctly, *then* he will know $\alpha \wedge \beta$. In our concrete case, let CE, CI, MP be the conjunction elimination rule, the conjunction introduction rule, and modus ponens, respectively, and let the symbol ";" denote the composition of actions. Then our theorem must be: $K_i(\alpha \wedge (\alpha \to \beta)) \to \langle CE_i; MP_i; CI_i\rangle K_i(\alpha \wedge \beta)$, and not $K_i(\alpha \wedge (\alpha \to \beta)) \to K_i(\alpha \wedge \beta)$ as in the standard modal approach.

[1] Instead of $K_i\alpha \to \langle R_i\rangle K_i\beta$ we could also introduce a binary operator $K_i\alpha\langle R_i\rangle K_i\beta$ with the interpretation "in a state where the agent i knows α, after the application of the rule R he may know β". However, the former notation is closer to that of dynamic logic, whereas the latter one does not offer any obvious advantage.

In general, suppose that β follows from α in some basis logic (which is accepted by the agent) and that the agent knows α. For explicit knowledge we cannot assume that the agent automatically knows β. Let a proof of β from α be given, where the axioms and inference rules used in the proof are R^1,...,R^n (in this order, where the same axiom or inference rule may occur at different places in the sequence.) Then, instead of the monotonicity rule in the standard modal approach we have the axiom: $K_i\alpha \rightarrow \langle R_i^1; \ldots; R_i^n \rangle K_i\beta$, where R_i^k is i's reasoning action of applying the inference rule R^k $(k = 1, \ldots, n)$. This axiom says that if the agent i performs the sequence of actions corresponding to the rules $R^1, \ldots, R^n$ (in this order) then he will know β under the given circumstances. Whether or not the agent can come to this conclusion depends crucially on his logical ability. In this way we see that the logical omniscience problem can be solved easily in a natural way: we can describe agents whose knowledge may or may not be closed under logical laws. On the other hand we can still say that the agent thinks rationally, that he is not logically ignorant. Theoretically he may produce all logical truths, and all logical consequences of his knowledge, but only if he is interested in doing so, if he has enough time and memory, et cetera.

In the above argumentation we have made an implicit assumption. We have assumed that all premises, once known by the agent, are still available after the agent performs a reasoning step. In the previous example, if the agent forgets the premise α immediately after using modus ponens, then he cannot apply the conjunction introduction rule to come to the conclusion $\alpha \wedge \beta$. Thus, we have to postulate that the agent does not forget what he previously knows after performing some reasoning action. This assumption can be formalized using persistence axioms for knowledge, for example, $K_i\alpha \rightarrow [R_i]K_i\alpha$.

Are such persistence axioms reasonable? Only under two conditions. First, the truth value of α should not change over time. If α becomes false after i's inference using rule R then it is not reasonable to postulate that i still knows α after the use of R. This point should be taken into account when we formally define the language of our logic. In particular, if our language contains temporal indexicals then sentences containing them cannot be regarded as persistent. Second, the truth value of α may not change through the agent's actions. This excludes formulae such that $\neg K_i\beta$: it is possible that agent i does not know β now, but will know it as a result of his reasoning. In general, a formula in which a knowledge operator occurs essentially negative (i.e., within the scope of an odd number of the negation sign) is not a suitable candidate for a persistent one. So, we may assume that persistent formulae are built up from objective formulae using conjunction, disjunction, and the knowledge operators only.

4.1.2 The abstract action of reasoning

In order to define systems of dynamic-epistemic logic formally we can fix a basis logic and then associate with each axiom schema and each inference rule an atomic action. The formal language is then defined over this set of atomic actions. The logic comprises all theorems of dynamic logic and the specific epistemic axioms discussed above.

However, there are some problems with this approach. First, there might be many different, but equivalent axiomatizations of the basis logic, so the choice of the basic actions must be arbitrary. Moreover, as the resulting dynamic-epistemic system con-

tains dynamic logic entirely, it becomes very complex and therefore difficult to be dealt with. Even more importantly, in most cases we do not need to care about what course of actions the agents just carried out; we are only interested in the result of the actions, so to speak. We only need to know that a certain agent has carried out some reasoning steps, and after that he gains certain new information.

This last point leads us to another approach. We introduce an auxiliary action F_i with the following intended reading: do any one of the atomic actions (we don't know what action;) repeat the non-deterministic choice finitely many times (at least once, but we don't know how many times!) The action F_i could be interpreted as a course of thought of the agent i. From the viewpoint of dynamic logic: if the set of all atomic actions associated with the agent i and his basis logic is a finite set $\{r_i^1, \ldots, r_i^n\}$, then F_i can be viewed as $(r_i^1 \cup r_i^2 \cup \ldots \cup r_i^n)^+$, where the symbols $\cup$ and $^+$ denote choice and non-zero iteration, respectively.[2] The choice of the symbols F_i is not accidental at all: in temporal logic it stands for the operator "Future". It turns out that our auxiliary action behaves in the same manner as the future operator of temporal logic: the operator $\langle F \rangle$ satisfies all the axioms for the minimal temporal logic $\mathbf{K}_t\mathbf{4}$. It is no surprise at all: we know that the minimal temporal logic can be embedded into dynamic logic, and one way to do this is to take the iteration of an action to interpret the future operator. The formal language in which our dynamic-epistemic logics are formulated is called $\mathcal{L}_N^{DE}$ and will be defined in the following section.

4.2 Dynamic epistemic logic

4.2.1 The language of dynamic-epistemic logic

Definition 13 (The language $\mathcal{L}_N^{DE}$) Let $Agt = \{1, \ldots, N\}$ be a set of N agents and let $\mathcal{L}_N^K$ be the language of epistemic logic as defined in definition 1. $\mathcal{L}_N^{DE}$ is the least set such that

1. $\mathcal{L}_N^K \subseteq \mathcal{L}_N^{DE}$
2. If $\alpha \in \mathcal{L}_N^{DE}$ then $\neg\alpha \in \mathcal{L}_N^{DE}$
3. If $\alpha \in \mathcal{L}_N^{DE}$ and $\beta \in \mathcal{L}_N^{DE}$ then $(\alpha \to \beta) \in \mathcal{L}_N^{DE}$
4. If $\alpha \in \mathcal{L}_N^{DE}$ then $\langle F_i \rangle \alpha \in \mathcal{L}_N^{DE}$

Conjunction and disjunction are defined as usual. $[F_i]\alpha$ abbreviates $\neg\langle F_i \rangle \neg\alpha$. The formula $\langle F_i \rangle \alpha$ is read: "α is true after some course of thought of i", $[F_i]\alpha$ means "α is true after any course of thought of i". (We could think of $\langle F_i \rangle$ and $[F_i]$ as the modalities "at some future times" and "at all future times" of temporal logic, but now time is subjective time, i.e., agent-dependent, generated by the agent's actions.) Note that we do not allow the operator $\langle F_i \rangle$ to occur inside the scope of any knowledge operator. The reason is that such expressions are indexicals: they contain temporal indexicals

[2] In dynamic logic another form of iteration is considered, viz. the one that allows for running a program zero time, denoted by *. But one can easily extend dynamic logic to include non-zero iteration as well.

like "later" or "always" implicitly. We want to exclude indexical expressions from our language because they require special treatment, which could be very involved and may obscure more important points.

Definition 14 (Knowledge-persistent formulae) The sublanguage $\mathcal{L}_N^{K+}$ of $\mathcal{L}_N^K$ is the smallest set of formulae from $\mathcal{L}_N^K$ which contains all objective formulae and is closed under the condition: if $\alpha, \beta \in \mathcal{L}_N^{K+}$ and $i \in Agt$ then $\{(\alpha \wedge \beta), (\alpha \vee \beta), K_i\alpha\} \subseteq \mathcal{L}_N^{K+}$.

4.2.2 Axioms for dynamic-epistemic logic

Let us discuss some potential candidates for dynamic-epistemic axioms. We shall examine the common modal axioms and see if their dynamic-epistemic counterparts are suitable for formalizing explicit knowledge.

As I have argued in the previous section, if β can be derived from the premises $\alpha_1, \ldots, \alpha_m$ by means of the inference rules $R_1, \ldots, R_n$, then the correct corresponding epistemic axiom should be $K_i\alpha_1 \wedge \ldots \wedge K_i\alpha_m \rightarrow \langle R_i^1; \ldots; R_i^n\rangle K_i\beta$. Translated into the language $\mathcal{L}_N^{DE}$, keeping in mind the intuitive reading of the operator $\langle F_i\rangle$, the axiom becomes $K_i\alpha_1 \wedge \ldots \wedge K_i\alpha_m \rightarrow \langle F_i\rangle K_i\beta$. In particular, the following formulae could be assumed as axioms for explicit knowledge:

- $K_i\alpha \wedge K_i(\alpha \rightarrow \beta) \rightarrow \langle F_i\rangle K_i\beta$
- $K_i\alpha \wedge K_i\beta \rightarrow \langle F_i\rangle K_i(\alpha \wedge \beta)$
- $K_i(\alpha \wedge \beta) \rightarrow \langle F_i\rangle K_i\alpha$
- $\langle F_i\rangle K_i(\alpha \vee \neg\alpha)$

I have also argued that some persistence postulates must be assumed in order to guarantee that all premises, once known by an agent, are still available after the agent performs a reasoning step. The idea that known sentences remain known after any course of thought of an agent i can be expressed through the axiom $K_i\alpha \rightarrow [F_i]K_i\alpha$, provided that α is persistent.

According to their intuitive interpretation, the dual operators $\langle F_i\rangle$ and $[F_i]$ must satisfy at least the postulates of $\mathbf{K}_t\mathbf{4}$, the minimal temporal logic of transitive time (see also appendix A), i.e., the (temporal) necessitation rule (from α to infer $[F_i]\alpha$) and the following two axioms should be assumed:

- $[F_i](\alpha \rightarrow \beta) \rightarrow ([F_i]\alpha \rightarrow [F_i]\beta)$
- $[F_i]\alpha \rightarrow [F_i][F_i]\alpha$

Can stronger principles be imposed on the operator $[F_i]$? Linearity does not seem reasonable: courses of thought can go to several different directions. There are certainly many ways to extend one's knowledge, e.g., by applying two different inference rules. We can imagine that an agent currently has a certain information state s_0 where two sentences α_1 and α_2 are implicitly available. After some course of thought he knows α_1 explicitly, and after some other course of thought he knows α_2. In this way two

different new information states s_1 and s_2 are possible: in one state (say s_1) the formula α_1 is (explicitly) known, but not α_2, and in the other one α_2 is known, but not α_1.

Consider now the information state s_1 where α_1 has been established. If the necessary conditions to establish α_2 are still available as in the original state s_0, then the sequence of reasoning steps leading to α_2 could be started at s_1, leading to a new, more complete information state s_3 where both α_1 and α_2 are known explicitly. The agent could also arrive at s_3 by starting the process of deriving α_1 from s_2. Thus, the principle of directedness seems attractive: any two developments originating from the same point will eventually be merged again. This principle corresponds to the axiom $\langle F_i\rangle[F_i]\alpha \to [F_i]\langle F_i\rangle\alpha$. In modal logic, this formula is known as schema **(G)**.

To distinguish genuine knowledge from belief, the axiom $K_i\alpha \to \alpha$ can be assumed. This Truth axiom seems unproblematic in dynamic-epistemic settings. As to consistency, two variants of the Consistency axiom are possible. The first is $\neg K_i(\alpha \wedge \neg\alpha)$, which says that i does not believe obvious contradictions. In normal modal logics, that formula is equivalent to the formulae $K_i\alpha \to \neg K_i\neg\alpha$. However, this needs not be true in the context of dynamic-epistemic logic, because agents may believe two sentences without believing their conjunction. These two consistency axioms seem to be acceptable rationality postulates.

Let us now examine how the ability of the agents to introspect their knowledge can be captured within our dynamic framework. An agent's action of introspection can be considered one of his basic reasoning actions[3]. Thus, we may view agent i's introspection action as one part of his abstract action F_i. Consider positive introspection first. Suppose that i knows α. Can we infer that he will know after introspecting his knowledge that he knows α? Not necessarily! We can assume that i will know that he *previously* knows α, but to support the inference that after his introspection action the agent knows that he knows α we need one more argument, namely that i's knowledge of α will not be changed through his reasoning actions. We have argued previously that such a persistence axiom is reasonable for a subclass of formulae. Thus, we have the following axiom of positive introspection, which corresponds to the schema **(4)** in modal epistemic logic: $K_i\alpha \to \langle F_i\rangle K_iK_i\alpha$, provided that α is persistent.

The same argumentation can be used to show that the candidate for the negative introspection axiom $\neg K_i\alpha \to \langle F_i\rangle K_i\neg K_i\alpha$ is not acceptable. It can happen that after a reasoning step the agent knows something what he did not know previously.

4.2.3 Systems of dynamic-epistemic logic

Now we go on to define axiomatic systems for reasoning about the dynamics of knowledge. We have three groups of axioms: the usual axioms of the propositional calculus, axioms for temporal logic, and axioms governing the interaction between knowledge and reasoning activities.

Definition 15 (The system $\mathbf{DEK}_N$) The logic $\mathbf{DEK}_N$ (Dynamic-Epistemic $\mathbf{K}_N$) has the following axiom schemata:

[3]One may ask how seriously one can take introspection as action. Well, it is true that introspection may differ from the "genuine" reasoning actions in some aspects. However, the differences are not quite significant. It seems reasonable to treat introspection as test of a certain kind, which is used by the agents to reason about their own mental state.

(PC1) $\alpha \to (\beta \to \alpha)$

(PC2) $(\alpha \to (\beta \to \gamma)) \to ((\alpha \to \beta) \to (\alpha \to \gamma))$

(PC3) $(\neg\beta \to \neg\alpha) \to (\alpha \to \beta)$

(TL1) $[F_i](\alpha \to \beta) \to ([F_i]\alpha \to [F_i]\beta)$

(TL2) $[F_i]\alpha \to [F_i][F_i]\alpha$

(DE1) $K_i\alpha \wedge K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$ (Closure under implication)

(DE2) $K_i\alpha \to [F_i]K_i\alpha$, provided that $\alpha \in \mathcal{L}_N^{K+}$ (Persistence)

(DE3) $\langle F_i\rangle K_i(\alpha \to (\beta \to \alpha))$

(DE4) $\langle F_i\rangle K_i((\alpha \to (\beta \to \gamma)) \to ((\alpha \to \beta) \to (\alpha \to \gamma)))$

(DE5) $\langle F_i\rangle K_i((\neg\beta \to \neg\alpha) \to (\alpha \to \beta))$

The rules of inference are:

(MP) From α and $\alpha \to \beta$ to infer β (Modus ponens).

(NEC$_t$) From α to infer $[F_i]\alpha$ (Temporal necessitation).

The axioms **(PC1) – (PC3)** together with the rule **(MP)** axiomatize completely the propositional calculus. Together with **(TL1)**, **(TL2)** and **(NEC$_t$)** they form a complete axiomatization of the minimal temporal logic of transitive time. The remaining axioms and inference rules describe the dynamics of knowledge. Axiom **(DE1)** says that the agents are capable of using modus ponens. Axiom **(DE2)** is the persistence axiom discussed previously, which says that agents do not forget what they know when they are reasoning. Axioms **(DE3) – (DE5)** state that the agents are able to use the axioms **(PC1) – (PC3)** of classical logic in their reasoning.

The notions of a proof, a theorem, and a consistent set of formulae (with respect to the logic $\mathbf{DEK}_N$) are defined in the usual way. The provability relation wrt $\mathbf{DEK}_N$ is denoted $\vdash_{\mathrm{DEK_N}}$ as usual. Moreover, we say that a formula $\alpha \in L_K$ is PC-provable, in symbol $\vdash_{PC} \alpha$, just in case α can be proved using only instances of the schemata **(PC1) – (PC3)** (in the sublanguage L_K) and modus ponens.

Of course, we can postulate that the agents can use further simple tautologies and inference rule in their reasoning. For example, we can include axioms such as $K_i\alpha \wedge K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$, or $\langle F_i\rangle K_i(\alpha \vee \neg\alpha)$. However, this is not necessary at all, because they can be proved, as we shall see later.

Extensions of $\mathbf{DEK}_N$ can be obtained by adding more axioms and inference rules to the basic system. We consider logics obtained from $\mathbf{DEK}_N$ by adding axioms from the following list:

(TL3) $\langle F_i\rangle[F_i]\alpha \to [F_i]\langle F_i\rangle\alpha$

(DE6) $K_i\alpha \to \alpha$

(DE7) $\langle F_i \rangle K_i (K_i \alpha \to \alpha)$

(DE8) $K_i \alpha \to \langle F_i \rangle K_i K_i \alpha$, provided that $\alpha \in \mathcal{L}_N^{K+}$

Axiom **(TL3)** corresponds to the directedness property discussed previously. It says that courses of thought are directed towards more epistemic completeness. Axiom **(DE6)** is the well-known schema **T** saying that knowledge entails truth. Axiom **(DE7)** says that agents potentially trust their knowledge: when thinking about themselves, they think that what they know must be true. Finally, **(DE8)** says that the agents are capable of positive introspection.

Definition 16 (Extensions of $\mathbf{DEK}_N$) Some extensions of the logic $\mathbf{DEK}_N$ are specified as follows:

- $\mathbf{DEK}_N^*$ is $\mathbf{DEK}_N$ plus **(TL3)**
- $\mathbf{DES4}_N$ is $\mathbf{DEK}_N$ plus **(DE6)**, **(DE7)** and **(DE8)**
- $\mathbf{DES4}_N^*$ is $\mathbf{DES4}_N$ plus **(TL3)**

The systems $\mathbf{DES4}_N$ and $\mathbf{DES4}_N^*$ can be viewed as logics of explicit, true knowledge. They correspond to the modal system $\mathbf{S4}_N$, as they require knowledge to be true, and the agents to have positive introspection. It is easy to see that both $\mathbf{DEK}_N^*$ and $\mathbf{DES4}_N$ contain $\mathbf{DEK}_N$ and are contained in $\mathbf{DES4}_N^*$, but neither is a subsystem of the other.

4.2.4 Some features of dynamic-epistemic logic

Theorem 17 (Consistency) The systems $\mathbf{DEK}_N$, $\mathbf{DEK}_N^*$, $\mathbf{DES4}_N$, and $\mathbf{DES4}_N^*$ are consistent.

Proof As $\mathbf{DEK}_N$, $\mathbf{DEK}_N^*$ and $\mathbf{DES4}_N$ are subsystems of $\mathbf{DES4}_N^*$, it suffices to show that $\mathbf{DES4}_N^*$ is consistent.

To see that the system $\mathbf{DES4}_N^*$ is consistent, i.e., no contradiction can be derived from it, it suffices to notice that all axioms and inference rules of $\mathbf{DES4}_N^*$ can be mapped to valid formulae and inference rules of the propositional calculus by deleting all occurrences of K_i and $\langle F_i \rangle$ from them. Therefore, all theorems of $\mathbf{DES4}_N^*$ must become propositional tautologies when all occurrences of K_i and $\langle F_i \rangle$ are deleted. Hence, a formula like $\alpha \wedge \neg\alpha$ cannot be derived.

The following theorem states that all the defined systems $\mathbf{DEK}_N$, $\mathbf{DEK}_N^*$, $\mathbf{DES4}_N$, and $\mathbf{DES4}_N^*$ solve the logical omniscience problem. It says that none of the rules **NEC**, **MON**, and **CGR** is valid. Moreover, an agent's explicit knowledge at a time, i.e., the totality of all what this agent knows at that time, needs not be closed under any nontrivial logical rule.

Theorem 18 (Non-Omniscience) 1. The following inference rules are not derivable in the systems $\mathbf{DEK}_N$, $\mathbf{DEK}_N^*$, $\mathbf{DES4}_N$, and $\mathbf{DES4}_N^*$:

(NEC) From α to infer $K_i\alpha$

(MON) From $\alpha \to \beta$ to infer $K_i\alpha \to K_i\beta$

(CGR) From $\alpha \leftrightarrow \beta$ to infer $K_i\alpha \leftrightarrow K_i\beta$

2. The following formulae are not provable in the systems $\mathbf{DEK}_N$, $\mathbf{DEK}^*_N$, $\mathbf{DES4}_N$, and $\mathbf{DES4}^*_N$:

   (a) $K_i(\alpha \to \beta) \to (K_i\alpha \to K_i\beta)$
   (b) $K_i\beta \to K_i(\alpha \to \beta)$
   (c) $K_i(\alpha \wedge \beta) \to K_i\alpha$
   (d) $K_i(\alpha \wedge \beta) \to K_i\alpha \wedge K_i\beta$
   (e) $K_i\alpha \wedge K_i\beta \to K_i(\alpha \wedge \beta)$
   (f) $K_i\alpha \to K_i(\alpha \vee \beta)$
   (g) $K_i\alpha \vee K_i\beta \to K_i(\alpha \vee \beta)$
   (h) $K_i\neg\neg\alpha \to K_i\alpha$
   (i) $K_i\alpha \to K_i\neg\neg\alpha$
   (j) $K_i\alpha \to K_iK_i\alpha$
   (k) $\neg K_i\alpha \to K_i\neg K_i\alpha$

Proof We can construct easily interpretations such that (i) all axioms of the dynamic-epistemic logic under consideration are valid, (ii) the rules of inference lead from valid formulae to valid ones, and (iii) the formulae and inference rules listed above are invalidated. We omit the details.

An agent described by the given logics is not logically omniscient. On the other hand, we cannot say that he is not rational: the agent *is* rational, because he can (at least in principle) perform actions to close his knowledge under logical laws, as the following theorems show. Instead of the necessitation rule and monotony rule in modal epistemic logic we have now a theorem stating that the agents *can* know all classical theorems and *can* draw all consequences of what they know, *provided that* they perform the right reasoning.

Theorem 19 Let α, β be objective formulae and let Λ be one of $\mathbf{DEK}_N$, $\mathbf{DEK}^*_N$, $\mathbf{DES4}_N$, and $\mathbf{DES4}^*_N$.

($\mathbf{NEC}_{de}$) If $\vdash_{PC} \alpha$ then $\vdash_\Lambda \langle F_i\rangle K_i\alpha$.

($\mathbf{MON}_{de}$) If $\vdash_{PC} \alpha \to \beta$ then $\vdash_\Lambda K_i\alpha \to \langle F_i\rangle K_i\beta$.

($\mathbf{MON}^*_{de}$) If $\vdash_{PC} \alpha \to \beta$ then $\vdash_\Lambda \langle F_i\rangle K_i\alpha \to \langle F_i\rangle K_i\beta$.

Proof First, note that $[F_i]\alpha \wedge \langle F_i\rangle\beta \to \langle F_i\rangle(\alpha \wedge \beta)$ and $\langle F_i\rangle\langle F_i\rangle\alpha \to \langle F_i\rangle\alpha$ are $\mathbf{K}_t\mathbf{4}$-provable and therefore $\mathbf{DEK}_N$-provable. Moreover, if $\alpha \to \beta$ is a theorem then so is $\langle F_i\rangle\alpha \to \langle F_i\rangle\beta$. We shall make extensive use of these facts in our proof without

mentioning them explicitly. To shorten the proofs we assume that all derivable formulae and rules of **PC** and $\mathbf{K}_t\mathbf{4}$ have been derived, so we do not have to write them down explicitly.

Consider rule $\mathbf{(NEC}_{de}\mathbf{)}$. Let $\vdash_{PC} \alpha$. We show $\vdash_\Lambda \langle F_i\rangle K_i\alpha$ by induction on the length m of the proof of α. If $m = 1$ then α must be an instance of one of the axiom schemata **(PC1)–(PC3)**. The claim follows from **(DE3)–(DE5)**. If $m > 1$ then α must be obtained by applying modus ponens from, say, β and $\beta \to \alpha$, which are PC-provable in less than m steps. So we assume that there is a PC-proof of α of length m where in the k-th and l-th lines we had proved β and $\beta \to \alpha$. The PC-proof of α can be extended to a $\mathrm{DES4_n}$-proof of $K_i\alpha$ as follows:

(k)	β	Assumption
(l)	$\beta \to \alpha$	Assumption
(m)	α	(k), (l), **(MP)**
(m+1)	$\langle F_i\rangle K_i\beta$	Ind. Hyp., (k)
(m+2)	$[F_i]\langle F_i\rangle K_i\beta$	(m+1), $\mathbf{(NEC}_t\mathbf{)}$
(m+3)	$\langle F_i\rangle K_i(\beta \to \alpha)$	Ind. Hyp., (l)
(m+4)	$K_i(\beta \to \alpha) \to [F_i]K_i(\beta \to \alpha)$	**(DE2)**
(m+5)	$\langle F_i\rangle[F_i]K_i(\beta \to \alpha)$	(m+3), (m+4)
(m+6)	$\langle F_i\rangle(\langle F_i\rangle K_i\beta \wedge [F_i]K_i(\beta \to \alpha))$	(m+2), (m+5)
(m+7)	$\langle F_i\rangle\langle F_i\rangle(K_i\beta \wedge K_i(\beta \to \alpha))$	(m+6)
(m+8)	$\langle F_i\rangle\langle F_i\rangle\langle F_i\rangle K_i\alpha$	(m+7), **(DE1)**
(m+9)	$\langle F_i\rangle K_i\alpha$	(m+8)

The rule $\mathbf{(MON}_{de}\mathbf{)}$ can now be derived as follows:

(1)	$\alpha \to \beta$	Assumption
(2)	$\langle F_i\rangle K_i(\alpha \to \beta)$	(1), $\mathbf{(NEC}_{de}\mathbf{)}$
(3)	$K_i\alpha \to [F_i]K_i\alpha$	**(DE2)**
(4)	$K_i\alpha \to ([F_i]K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta))$	(2), (3)
(5)	$K_i\alpha \to \langle F_i\rangle(K_i(\alpha \to \beta) \wedge K_i\alpha)$	(4)
(6)	$K_i\alpha \to \langle F_i\rangle\langle F_i\rangle K_i\beta$	(5), **(DE1)**
(7)	$K_i\alpha \to \langle F_i\rangle K_i\beta$	(6)

To prove $\mathbf{(MON}^*_{de}\mathbf{)}$ we apply $\mathbf{(MON}_{de}\mathbf{)}$ to derive $K_i\alpha \to \langle F_i\rangle K_i\beta$ from $\alpha \to \beta$. Then a rule of $\mathbf{K}_t\mathbf{4}$ can be used to infer $\langle F_i\rangle K_i\alpha \to \langle F_i\rangle\langle F_i\rangle K_i\beta$. Using the $\mathbf{K}_t\mathbf{4}$-theorem $\langle F_i\rangle\langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i\beta$ we get the desired result.

Corollary 20 Assume that α, β are objective formulae. The following formulae are theorems of $\mathbf{DEK}_N$ and its extensions:

1. $K_i\beta \to \langle F_i\rangle K_i(\alpha \to \beta)$
2. $K_i(\alpha \wedge \beta) \to \langle F_i\rangle K_i\alpha$
3. $K_i(\alpha \wedge \beta) \to \langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i\beta$
4. $K_i\alpha \to \langle F_i\rangle K_i(\alpha \vee \beta)$

5. $K_i\alpha \vee K_i\beta \rightarrow \langle F_i\rangle K_i(\alpha \vee \beta)$
6. $K_i\neg\neg\alpha \rightarrow \langle F_i\rangle K_i\alpha$
7. $\langle F_i\rangle K_i\beta \rightarrow \langle F_i\rangle K_i(\alpha \rightarrow \beta)$
8. $\langle F_i\rangle K_i(\alpha \wedge \beta) \rightarrow \langle F_i\rangle K_i\alpha$
9. $\langle F_i\rangle K_i\alpha \rightarrow \langle F_i\rangle K_i(\alpha \vee \beta)$

Probably, the above rules and theorems are derivable for a larger class of formulae, not only for objective ones. The following list comprises some more provable formulae of $\mathbf{DEK}_N$ and its extensions. They say that if all premises of a valid inference rule are known, then after some steps of reasoning the agent will know the conclusion. This still holds if one of the premises is not known yet but will be known after some reasoning. The theorem is proved in appendix B.

Theorem 21 Assume that α and β are objective. The following formulae are theorems of $\mathbf{DEK}_N$ and its extensions:

1. $K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \rightarrow \beta) \rightarrow \langle F_i\rangle K_i\beta$
2. $K_i\alpha \wedge \langle F_i\rangle K_i\beta \rightarrow \langle F_i\rangle K_i(\alpha \wedge \beta)$
3. $K_i\alpha \wedge K_i\beta \rightarrow \langle F_i\rangle K_i(\alpha \wedge \beta)$
4. $K_i(\alpha \wedge \beta) \rightarrow \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$
5. $\langle F_i\rangle K_i(\alpha \wedge \beta) \rightarrow \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$

4.2.5 Systems with the directedness axiom

The following theorem states some results for specific systems which will clarify the role played by the directedness axiom **(TL3)**. Observe that the formula $\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \rightarrow \beta) \rightarrow \langle F_i\rangle K_i\beta$ is not provable in $\mathbf{DEK}_N$, i.e., it may be the case that both $\langle F_i\rangle K_i\alpha$ and $\langle F_i\rangle K_i(\alpha \rightarrow \beta)$ are true but $\langle F_i\rangle K_i\beta$ is not true. Generally, if a valid inference rule has at least 2 premises, and if each of these premises will be known after some course of thought, then it is not necessarily the case that the conclusion will be known. Such situations are precluded in the presence of the directedness axiom.

Theorem 22 Let α and β be objective formulae. In logics containing the schema **(TL3)**, the following formulae are provable:

1. $\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \rightarrow \beta) \rightarrow \langle F_i\rangle K_i\beta$
2. $\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i\beta \rightarrow \langle F_i\rangle K_i(\alpha \wedge \beta)$

Proof See appendix B.

Utilizing the previous result we can establish an embedding relation between $\mathbf{K}_N$ and $\mathbf{DEK}_N^*$. Similar relations obtain between other normal modal systems and their dynamic-epistemic counterparts which contain schema **(TL3)**.

Theorem 23 Let $\alpha \in \mathcal{L}_N^K$ be a formula whose modal depth is at most 1. Let α' be the $\mathcal{L}_N^{DE}$-formula obtained by replacing every occurrence of K_i in α by $\langle F_i \rangle K_i$. Then α is a theorem of $\mathbf{K}_N$ if and only if α' is a theorem of $\mathbf{DEK}_N^*$.

Proof Let α be a $\mathbf{K}_N$-theorem. We show by induction over the proof length that α' is a $\mathbf{DEK}_N^*$-theorem. If α is a propositional tautology then α' is also a propositional tautology. If α is an instance of the schema **(K)** then α' is an instance of the schema $\langle F_i \rangle K_i \alpha \wedge \langle F_i \rangle K_i (\alpha \rightarrow \beta) \rightarrow \langle F_i \rangle K_i \beta$, which is a $\mathbf{DEK}_N^*$-theorem according to theorem 22. If α has been derived from β and $\beta \rightarrow \alpha$ by means of modus ponens then β' and $(\beta \rightarrow \alpha)'$ are both theorems of $\mathbf{DEK}_N^*$, by induction hypothesis. It can be easily seen that $(\beta \rightarrow \alpha)'$ is the same formula as $\beta' \rightarrow \alpha'$. So applying modus ponens yields that α' is a theorems of $\mathbf{DEK}_N^*$. Finally, suppose that α has been derived from β using the knowledge necessitation rule **(NEC)**. Then α is $K_i \beta$ and α' is $\langle F_i \rangle K_i \beta$. Because α hat at most the modal depth 1, β must be objective, so $\langle F_i \rangle K_i \beta$ can be derived according to theorem 19.

To prove the converse, observe that every $\mathbf{DEK}_N^*$-proof can be transformed to a $\mathbf{K}_N$-proof by deleting all occurrences of $\langle F_i \rangle$. (Recall that $[F_i]\alpha$ is just an abbreviation for $\neg \langle F_i \rangle \neg \alpha$). It can also be easily verified that if α' is obtained from α by replacing every occurrence of K_i by $\langle F_i \rangle K_i$, then deleting all occurrences of $\langle F_i \rangle$ in α' yields α again. Therefore, if α' is $\mathbf{DEK}_N^*$-derivable then α is $\mathbf{K}_N$-derivable.

Algorithmic knowledge

In the previous chapter I have developed a framework for reasoning about explicit knowledge. The strategy is to take the cost of inferring new information into account. Following this strategy a number of logics have been defined which can solve all variants of the logical omniscience problem and at the same time can account for the intuition that agents are rational beings. In my framework it is possible to model situations where an agent's explicit knowledge is not closed under any logical law: he may know all premises of an inference rule without knowing the conclusion. But this does not mean he is logically ignorant. On the contrary, he may well be perfectly rational: if he chooses to draw a conclusion of his knowledge and if he has sufficient computational resources, he will eventually succeed in doing it. Thus, resource-bounded reasoning can be modeled realistically: an agent's lack of logical omniscience stems from his resource-boundedness, and not from his inability to use certain logical rules.

However, there are a number of situations where resource-bounded reasoning cannot be modeled within the framework of explicit knowledge considered so far. First of all, the dynamic-epistemic systems of the previous chapter are based on standard qualitative temporal logic and are therefore not suited to describe quantitative time constraints[1]. Moreover, they have too little expressive power for modeling meta-reasoning, i.e., for modeling how an agent reasons about the reasoning process of himself or of other agents.

In this chapter I shall introduce a new concept of knowledge which allows quantitative resource constraints to be formalized directly. This concept generalizes both concepts of explicit and implicit knowledge and avoids the problems of the existing approaches. In the next section I shall discuss the concept informally. Then I proceed to define a formal language and some formal systems for resource-bounded reasoning about knowledge. Finally, comparisons with other notions of knowledge will be made.

[1]Among the relevant resources, time is the most important one, so we shall focus on that factor and try to model time constraints. The other resources can be modeled in a similar manner, as I shall show later.

5.1 Motivation

5.1.1 Why explicit knowledge is not enough

What an intelligent agent chooses to do depends on his available resources. The available resources are typically measured in quantitative, rather than qualitative terms. A logic based on a qualitative time structure can model qualitative constraints like "agent 1 always knows α before agent 2 knows it". It fails, however, when the constraints placed on the resources are given in quantitative terms. It is not possible to express, e.g., that an agent can compute the solution to some problem within a certain time period.

The language $\mathcal{L}_N^{DE}$ of dynamic-epistemic logic does not allow the operator $\langle F_i \rangle$ to occur within the scope of any knowledge operator. Consequently, the capacities of the dynamic-epistemic systems to model meta-reasoning are rather restricted. For example, $K_1 K_2 \alpha$ ("agent 1 knows that agent 2 knows α") and $\langle F_1 \rangle K_1 K_2 \alpha$ ("after some reasoning agent 1 will know that agent 2 knows α") are well-formed formulae, but $K_1 \langle F_2 \rangle K_2 \alpha$ ("agent 1 knows that agent 2 will know α after some reasoning") and $\langle F_1 \rangle K_1 \langle F_2 \rangle K_2 \alpha$ ("after some reasoning agent 1 will know that agent 2 will know α after some reasoning") are not.

To be able to model resource boundedness within the language we consider another notion of knowledge. The main intuition is the following. An agent's action depends not only on what he *currently* knows, but also on what he is able to infer within some specific amount of time (intuitively, the time within which a decision must be made — a classical example being the time available to make the next move in chess.) Given that an agent needs to accomplish a task within an hour but does not yet know what actions he must perform, it cannot be inferred that he will not finish his job in time: he may be able to calculate the plan and finish it before the deadline. If an agent knows that another agent must act under some time constraint, he may infer what the second agent can or cannot know under this constraint and predict his action accordingly. Therefore, it must be considered what agents can reliably know within $1, 2, 3, \ldots$ time units, and not only what they currently know, i.e., what they know within 0 unit of time. Thus, we shall analyze sentences of the form "if asked about α, i can derive reliably the answer within n time units", instead of sentences of the form "agent i knows α (now)".

In the informal characterization of knowledge above, the qualification "reliably" is important. It distinguishes an agent's ability to bring about certain states of affairs from the mere logical possibility that such states of affairs may obtain. The difference can be illustrated by an example. A shooter may accidentally hit a target at 1 km distance, but it cannot be said that he has that ability. It cannot be safely assumed that he will succeed if he decides to try. He may hit the target once but the success is not repeatable. Hence, although there is the possibility that he can perform a certain action, he does not have the ability to do it. In the context of reasoning actions, an agent may possess a large number of algorithms which can be applied to compute knowledge. If he chooses to derive α from his current knowledge, he may by chance succeed very quickly if he applies the right algorithm. However, if he happens to select another algorithm then it may take very long to compute the same sentence. It can

even the case that the algorithm does not terminate at all. But if it cannot be safely assumed that the agent can compute α in time, then generally the possible knowledge of α is not enough to justify his action. Reliability implies that the agent is able to select deterministically a suitable procedure for the input and compute the answer within finite time.

Our goal is to represent not only *what* agents know or can know, but also *when* they are expected to know what they can know. The first question is answered by specifying the logic used by agents in their reasoning, and the second one by a complexity analysis. What an agent knows or can derive from his knowledge is determined by the logic he uses in his reasoning. An agent may not know a sentence *now*, but he may possess a procedure to infer that sentence and know it at some future point. The amount of time needed to compute that knowledge depends on several factors, of which the most important ones are the complexity of the sentence and the agent's reasoning power. If the complexity of a sentence and the computation speed of an agent are known then the time he needs to infer the sentence can be estimated.

5.1.2 The language of algorithmic knowledge

For modeling knowledge with time constraints we need to use some model of time measurement. As remarked previously, we shall deal with sentences of the form "if asked about α, agent i can derive reliably the answer within n time units". For simplicity we shall use natural numbers to measure time, i.e., we assume that n is a natural number. So the language we consider should contain formulae of the form $K_i^n\alpha$ where i is the name of an agent, n is a natural number, and α is a formula. The formula $K_i^n\alpha$ can be read "agent i knows α within n units of time" and is interpreted: "if agent i chooses to derive α from his current knowledge, then after at most n time units he will succeed", or alternatively, "if asked about α, i is able to derive reliably the answer 'yes' within n units of time". That is, we require not only that agent i have at least one procedure to compute α, but also that i be able to choose the correct procedure leading to α under the given time constraint, namely, to arrive at the conclusion α after at most n time units[2].

Sometimes it can be assumed safely that an agent is able to infer some fact, but it is not possible to estimate accurately how long the computation would take. For example, the complexity of the employed algorithm or the agent's inference strategy may not be known completely. To deal with such cases we introduce a sort of quantification into the language. We consider statements of the form "there is a number x such that the agent i is able to compute the fact α within x units of time". Such a statement is formalized by the formula $K_i^{\exists}\alpha$, which can be read: "agent i can infer α reliably in finite time". That is, when presented with the fact α, the agent is able to choose a suitable algorithm which runs on α and terminates with the (correct) answer after finitely many steps.

The formula $K_i^n\alpha$ entails the following facts about the agent's i information state.

[2]The sentence "agent i needs n time units to compute α" does not imply that i *will* know α at time $t_{now}+n$, where t_{now} is the current time. If the agent is not asked to provide the information α, then he has no reason to waste his resources in order to find a useless answer. The aspect of goal-directedness is implicit in our concept of knowledge.

First, the formula α follows (with respect to the logic used by i) from all what i knows. Second, agent i has an algorithm to establish that connection and which he is able to select to use when he chooses to compute α. Third, that computation takes at most n time unit. The formula $K_i^{\exists}\alpha$ is weaker in the sense that it does not tell exactly how long the computation of α will take. It only says that the computation is guaranteed to terminate.

Our notion of knowledge can be called algorithmic knowledge: knowledge is tied up with an algorithm to establish it. It represents not only factual knowledge but also a kind of procedural knowledge. The concepts of explicit and implicit knowledge can be regarded naturally as two special cases of algorithmic knowledge. Explicit knowledge can be defined as $K_i^0\alpha$, which says that agent i has immediate access to the information α and can act on it. Implicit knowledge is defined as $K_i^{\exists}\alpha$: agent i knows α implicitly if he is able to compute α when required. This is, however, not the only way to define a useful notion of implicit knowledge. For instance, one can stipulate that an agent knows a fact implicitly it can be inferred from his explicit knowledge (with respect to some inference system).

Our use of the term "algorithmic knowledge" as explained above should not be confused with other usages found elsewhere in the literature. Binmore and Shin ([BS92]) use the term to emphasize the connection between knowledge and provability (see also [SW94]). In their terminology, an agent's algorithmic knowledge is whatever the agent can infer using a Turing machine. The properties of this concept are studied and related to properties of provability concepts. Halpern, Moses, and Vardi ([HMV94]) also define algorithmic knowledge in terms of computation: an agent is said to know a fact at a certain state if at that state he can compute that he knows that fact. That is, given his local data, his local algorithm terminates and outputs the answer "Yes" when presented with the fact. Clearly, these concepts describe knowledge that is not necessarily available immediately to the agent. They are in spirit related to our concept of implicit knowledge, defined above as $K_i^{\exists}\alpha$. Hence, they both fall under the category of implicit knowledge in our classification of chapter 2: they characterize a kind of information that is implicitly available to an agent but must be computed and made explicit before the agent can act upon.

Formally, the language $\mathcal{L}_N^{AK}$ of algorithmic knowledge for N agents is defined as follows:

Definition 24 Let ω be the set of natural numbers, $Agent = \{1, \ldots, N\}$ a set of agents and $Atom$ a countable set of atomic formulae. The set of formulae is the least set $\mathcal{L}_N^{AK}$ such that

- $Atom \subseteq \mathcal{L}_N^{AK}$
- If $\alpha \in \mathcal{L}_N^{AK}$ then $\neg\alpha \in \mathcal{L}_N^{AK}$
- If $\alpha \in \mathcal{L}_N^{AK}$ and $\beta \in \mathcal{L}_N^{AK}$ then $(\alpha \to \beta) \in \mathcal{L}_N^{AK}$
- If $i \in Agent$, $n \in \omega$, and $\alpha \in \mathcal{L}_N^{AK}$ then $K_i^n\alpha \in \mathcal{L}_N^{AK}$
- If $i \in Agent$ and $\alpha \in \mathcal{L}_N^{AK}$ then $K_i^{\exists}\alpha \in \mathcal{L}_N^{AK}$

The rationality of agents is expressed through two capacities: first, the ability to draw logical consequences from what is already known, and second, the ability to compute the complexities of certain reasoning problems in order to infer when something can be known. It should be stressed that these two capacities are implementable. Agents have been frequently supplied with inference machines which allow them to infer new information from what has been known. The complexities of many problems can be computed at a low cost from their syntactic structures alone, so it is not hard to build into agents the capability to recognize the structure of a problem and estimate the cost to solve it. It turns out that we can develop quite rich theories of the algorithmic notion of knowledge we have introduced. To develop logics of algorithmic knowledge we try to establish logical relationships among the formulae of the language $\mathcal{L}_N^{AK}$. This is done by developing the framework for reasoning about explicit knowledge (chapter 4) a step further.

5.2 Reasoning about algorithmic knowledge

Our logics of algorithmic knowledge will be built up step by step from some basis logic. We shall take the propositional calculus as the basis and develop epistemic systems by adding (proper) epistemic laws to this basis. Now let us see how such laws may look like. We make the simplifying assumption that all agents have the same formal language and employ the same logic in their reasoning.

5.2.1 Axioms for algorithmic knowledge

Let us assume that an agent i knows α within m units of time, i.e., he needs m time units to infer α. Then naturally he is able to do it when even more time is available. So we can take as axiom any formula $K_i^m\alpha \rightarrow K_i^n\alpha$ where $m < n$. Note that this axiom *does not* say that knowledge is persistent in the sense that once established it will be available henceforth. The formula $K_i^m\alpha$ does not entails that i will know α at time point m. It does not even imply that α will eventually be known at all. In this aspect the present approach makes a more realistic assumption than the persistence axiom in chapter 4.

We have remarked previously that the formula $K_i^n\alpha$ contains more information than $K_i^{\exists}\alpha$. While the latter formula only says that agent i is able to derive α in finite time if he chooses to, the former one also specifies the amount of time needed by i to complete that task. Thus, the implication $K_i^n\alpha \rightarrow K_i^{\exists}\alpha$ can be assumed as an axiom.

Let α be a provable formula of the logic used by agent i. We have argued previously that it cannot be assumed that i knows α automatically (i.e., without any reasoning.) However, he may know it after some course of reasoning. The interesting question is whether or not he is able to compute α *reliably* within finite time. That would be the case if the agent has a general strategy which fulfills the following two requirements. First, it selects for any formula of the language an algorithm to compute that formula. Second, if the formula is provable then the selected algorithm will terminate with the correct answer. I shall argue that under reasonably weak assumptions, such a strategy exists and can be adopted by any intelligent agent, so that $K_i^{\exists}\alpha$ can be safely postulated if α is provable.

The set of axioms that any agent presupposes is decidable — in the normal case even finite. Because the number of permissible inference rules is also limited, all proofs can be generated algorithmically. Hence, there is a general-purpose theorem prover that can validate any theorem in finitely many steps. If the agent's selection mechanism always returns that general algorithm for computing knowledge, he is able to validate every theorem α. That is, when presented with a theorem α he can select an algorithm which runs on α and outputs the answer "Yes" after a finite number of steps. Although the described strategy ("always use the same algorithm") satisfies the stated conditions, it may not be the best: specific problems may be solved much more quickly by special algorithms than by a general-purpose theorem prover. Hence, the following would be a more reasonable strategy. First, the agent analyzes the query α and tries to select one of his special algorithms to infer it. If no such algorithm can be found, then the general algorithm is selected. In this way, he can always find an algorithm to verify α. (If the selection mechanism is not reliable, i.e., if it could return a wrong algorithm for some queries, then several algorithms can be selected and executed concurrently or interleavingly.)

A strategy to successfully prove every provable formula can be acquired by rational agents. An intelligent agent may learn to use some algorithms to compute knowledge. Such algorithms (together with a suitable selection scheme) can be built into artificial agents. Hence, the rule to infer $K_i^{\exists}\alpha$ from α can be assumed to be valid.

The statement $K_i^{\exists}\alpha$ contains some uncertainty. It is not clear how long agent i will need to infer α. Can a more definite statement be made? That is, can a natural number n (which typically depends on the structure of the theorem α) be determined such that $K_i^n\alpha$ can be assumed as a postulate? The discussion of this question will be postponed until section 5.2.3.

Now suppose that $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ is provable and that each of the formulae $\alpha_1, \ldots, \alpha_n$ can be computed reliably by an agent i in finite time, i.e., $K_i^{\exists}\alpha_1, \ldots, K_i^{\exists}\alpha_n$ are regarded to be true. Is it reasonable to infer that i can compute β reliably if he chooses to derive it? I shall argue that the conclusion $K_i^{\exists}\beta$ can be justified.

When presented with a question β, an agent i naturally attempts to derive β from all what he knows[3]. It is reasonable to assume that the consequence relation used by a rational agent has a certain transitivity property: if all the formulae $\alpha_1, \ldots, \alpha_n$ are derivable from some knowledge base and β can be inferred from $\alpha_1, \ldots, \alpha_n$, then β can also be inferred from that knowledge base. Thus we can assume that β follows from all what i knows. Because agents are assumed to process only a limited amount of information, every consequence of their knowledge can be computed algorithmically. With a suitable selection strategy, e.g., one of the strategies outlined previously, an agent can find an algorithm to compute his knowledge successfully. Consequently, agent i is able to compute β after a finite number of steps. So, if $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ is a theorem then we can assume that $K_i^{\exists}\alpha_1 \wedge \ldots \wedge K_i^{\exists}\alpha_n \rightarrow K_i^{\exists}\beta$ is valid.

As a special case we can assume that $K_i^{\exists}\alpha \wedge K_i^{\exists}(\alpha \rightarrow \beta) \rightarrow K_i^{\exists}\beta$ is valid, which says that agent i can use modus ponens in his reasoning: if he can derive both α and $\alpha \rightarrow \beta$

[3] An agent may in fact have some relevance criteria to narrow down the search space, so he actually tries to infer β from the relevant part of his knowledge. However, it is typically not possible to restrict the attention to the formula $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$, because the knowledge that β can be derived from the intermediate results $\alpha_1, \ldots, \alpha_n$ can usually be obtained only after a proof has been constructed.

then he is also able to derive β. Because explicit knowledge implies implicit knowledge, the formula $K_i^0\alpha_1 \wedge \ldots \wedge K_i^0\alpha_n \rightarrow K_i^\exists\beta$ is valid, provided that $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ is a theorem. Thus, agents are able to compute all logical consequences of their explicit knowledge.

Recall that in chapter 4, the persistence axiom ("everything that has been once deduced will be available henceforth") plays a prominent role in justifying the postulates stating that agents are able to use logical laws in their reasoning. In the current approach such an assumption is not necessary because we argue at a higher abstraction level. Only the final result is important, not the intermediate ones.

Since the formula $K_i^\exists\alpha \rightarrow \alpha$ is merely a definitional stipulation, it seems uncontroversial. If that axiom is adopted, the formula $K_i^n\alpha \rightarrow \alpha$ can be proved for all natural numbers n. More interesting are variants of the consistency axiom. A weak consistency criterion is that agents do not believe apparent contradictions, i.e., their explicit knowledge is consistent: $K_i^0\alpha \rightarrow \neg K_i^0\neg\alpha$. A stronger requirement is that an agent's implicit knowledge be consistent, which is captured by the schema $K_i^\exists\alpha \rightarrow \neg K_i^\exists\neg\alpha$. That is, the agent's explicit knowledge is free of contradictions and his inference procedures are sound, so that consistency is preserved. (The latter formula is indeed a stronger condition than the former one because $K_i^\exists\alpha$ follows from $K_i^0\alpha$.)

What about self-introspection? If an agent knows or does not know something explicitly, he only needs to reflect about himself to find it out. The cost of reflection is usually low, so it can be assumed that self-reflection can be performed in constant time. Hence, the formulae $K_i^0\alpha \rightarrow K_i^1K_i^0\alpha$ and $\neg K_i^0\alpha \rightarrow K_i^1\neg K_i^0\alpha$ seem plausible.

The situation is quite different when an agent i tries to compute the truth value of $K_i^n\alpha$ where $n > 0$. Although he may actually compute α and reflect about that, the result of his computation does not say much about his *ability* to infer α. He may succeed to compute α within n time units, but there is still a chance that this success is only accidental and not reproducible. On the other hand, even if $K_i^n\alpha$ is false (i.e., agent i cannot compute α reliably within n time units), i may still happen to successfully infer α after less than n time units. Consequently, it is not sound to infer that he can or cannot compute α reliably within n time units. So, generally neither $K_i^n\alpha \rightarrow K_i^{n+1}K_i^n\alpha$ nor $\neg K_i^n\alpha \rightarrow K_i^{n+1}\neg K_i^n\alpha$ is valid. The same argument applies to the formulae $K_i^\exists\alpha \rightarrow K_i^\exists K_i^\exists\alpha$ and $\neg K_i^\exists\alpha \rightarrow K_i^\exists\neg K_i^\exists\alpha$. Those principles can only be assumed for agents who know their abilities well. They can, for example, be postulated for an agent who works deterministically. For such an agent, a small number of tests may suffice to determine if he can perform a task under certain conditions.

5.2.2 Logics of algorithmic knowledge

The basic logic of algorithmic knowledge with N agents will be called $\mathbf{K}_N^A$. It is specified by the following axioms and rules of inference.

Definition 25 The logic $\mathbf{K}_N^A$ consists of the following axiom schemata and rules of inference:

(PC) All propositional tautologies

($\mathbf{K}^A$) $K_i^\exists\alpha \wedge K_i^\exists(\alpha \rightarrow \beta) \rightarrow K_i^\exists\beta$

(P^A) $K_i^m\alpha \to K_i^n\alpha$, for all $m, n \in \omega$ such that $m < n$

(Q^A) $K_i^n\alpha \to K_i^\exists\alpha$, for all $n \in \omega$

(MP) From α and $\alpha \to \beta$ to infer β (Modus ponens)

(NEC^A) From α to infer $K_i^\exists\alpha$

The definition of $\mathbf{K}_N^A$ calls for some explanation and comment. Axiom **(K^A)** and rule **(NEC^A)** correspond to the familiar modal postulates. However, the intended interpretation of the operator $K_i^\exists$ is now different: unlike the necessity operator $\Box$ of modal logic, which has an universal flavor ("true in all possible worlds"), the operator $K_i^\exists$ has a rather existential flavor ("the computation eventually terminates with the correct answer"). Hence, our postulates must be justified in a different way. The axiom schemata **(P^A)** and **(Q^A)** characterize the operators $K_i^n\alpha$ for natural numbers n. By means of **(Q^A)**, formulae like $K_i^0\alpha \wedge K_i^0(\alpha \to \beta) \to K_i^\exists\beta$ ("agents are able to compute consequences of their explicit knowledge") can be proved and need not be postulated separately. We do not have any axiom of the form $K_i^0\alpha$ because nothing can be assumed to be (explicitly) known a priori. However, for certain formulae α a number $n > 0$ can be determined such that $K_i^n\alpha$ may be assumed to be logically valid. We shall investigate such formulae later and use them to define more powerful logics of algorithmic knowledge.

Another way to enrich the basic system is to use postulates which capture additional properties of knowledge. We have discussed axioms which have often been used in the context of modal epistemic logic.

(T^A) $K_i^\exists\alpha \to \alpha$ (Truth axiom)

(D^A) $K_i^\exists\alpha \to \neg K_i^\exists\neg\alpha$ (Consistency axiom)

(4^A) $K_i^\exists\alpha \to K_i^\exists K_i^\exists\alpha$ (Positive introspection axiom)

(5^A) $\neg K_i^\exists\alpha \to K_i^\exists\neg K_i^\exists\alpha$ (Negative introspection axiom)

Adding suitable postulates from that list to the basic system will yield stronger logics of algorithmic knowledge. Those extensions of $\mathbf{K}_N^A$ are named in the same manner as the modal systems in chapter 2. For example, $\mathbf{S5}_N^A$ is the logic $\mathbf{K}_N^A$ plus the axioms $\mathbf{T}^A$, $\mathbf{4}^A$ and $\mathbf{5}^A$.

Theorem 26 (Consistency) $\mathbf{S5}_N^A$ and its subsystems are consistent.

Proof We map formulae of the language $\mathcal{L}_N^{AK}$ to propositional formulae by deleting all occurrences of the knowledge operators $K_i^\exists$ and K_i^n (for all $n \in \omega$) from them. By that transformation, all axioms of $\mathbf{S5}_N^A$ are mapped to propositional tautologies. Moreover, applying the resulting inference rules to propositional tautologies results in tautologies. Therefore, all $\mathbf{S5}_N^A$-theorems of become propositional tautologies. So, a contradiction like $\alpha \wedge \neg\alpha$ cannot be derived.

Obviously, $\mathbf{K}_N^A$ solves all variants of the logical omniscience problem with respect to the explicit concept of knowledge. To see that, it suffices to observe that the set $\{\neg K_i^0\alpha | \alpha \in \mathcal{L}_N^{AK}\}$ is consistent with $\mathbf{K}_N^A$, i.e., $\mathbf{K}_N^A$ can describe agents who (at some of their information states) know nothing explicitly. (However, they always know something implicitly.) Likewise, it is easy to see that what an agent explicitly knows (i.e., what he knows in 0 unit of time) needs not be closed under logical consequences or even under any logical law, e.g., $K_i^0\alpha \wedge K_i^0(\alpha \to \beta) \wedge \neg K_i^0\beta$ is perfectly $\mathbf{K}_N^A$-consistent. Moreover, $K_i^n\alpha \wedge K_i^n(\alpha \to \beta) \wedge \neg K_i^n\beta$ can also be seen to be consistent for any n. On the other hand, many closure properties hold for the notion of implicit knowledge. For example, $K_i^0\alpha \wedge K_i^0(\alpha \to \beta) \to K_i^\exists\beta$ is provable in $\mathbf{K}_N^A$. In general, agents described by our logic are rational in the sense that they can draw all logical consequences of their knowledge if the necessary resources are available, as the following lemma shows.

Lemma 27 The following rules of inference are valid for $\mathbf{K}_N^A$ and its extensions:

(MON^A) From $\alpha \to \beta$ to infer $K_i^\exists\alpha \to K_i^\exists\beta$

(CGR^A) From $\alpha \leftrightarrow \beta$ to infer $K_i^\exists\alpha \leftrightarrow K_i^\exists\beta$

Proof The rule **(CGR^A)** is a trivial consequence of **(MON^A)**. To prove **(MON^A)** let us suppose that $\alpha \to \beta$ is a theorem. By **(NEC^A)** we can infer $K_i^\exists(\alpha \to \beta)$. The formula $K_i^\exists(\alpha \to \beta) \to (K_i^\exists\alpha \to K_i^\exists\beta)$ is equivalent to **(K^A)** and is therefore a theorem of $\mathbf{K}_N^A$. So, $K_i^\exists\alpha \to K_i^\exists\beta$ can be inferred using modus ponens.

The next theorem shows that the common systems of modal epistemic logic can be embedded into the corresponding systems of logic for algorithmic knowledge. For that purpose we map each formula of the language $\mathcal{L}_N^K$ to a formula of $\mathcal{L}_N^{AK}$ and show that provability is preserved.

Theorem 28 Let the translation function $tr : \mathcal{L}_N^K \mapsto \mathcal{L}_N^{AK}$ be defined as follows:

- $tr(\phi) = \phi$ for all $\phi \in Atom$
- $tr(\neg\alpha) = \neg tr(\alpha)$
- $tr(\alpha \to \beta) = (tr(\alpha) \to tr(\beta))$
- $tr(K_i\alpha) = K_i^\exists tr(\alpha)$

A formula $\alpha \in \mathcal{L}_N^K$ is a theorem of a modal epistemic logic if and only if the $\mathcal{L}_N^K$-formula $tr(\alpha)$ is a theorem of the corresponding logic of algorithmic knowledge. Concretely:

1. $\vdash_{K_N} \alpha$ iff $\vdash_{K_N^A} tr(\alpha)$
2. $\vdash_{T_N} \alpha$ iff $\vdash_{T_N^A} tr(\alpha)$
3. $\vdash_{S4_N} \alpha$ iff $\vdash_{S4_N^A} tr(\alpha)$
4. $\vdash_{S5_N} \alpha$ iff $\vdash_{S5_N^A} tr(\alpha)$

5. $\vdash_{KD_N} \alpha$ iff $\vdash_{KD_N^A} tr(\alpha)$

6. $\vdash_{KD4_N} \alpha$ iff $\vdash_{KD4_N^A} tr(\alpha)$

7. $\vdash_{KD45_N} \alpha$ iff $\vdash_{KD45_N^A} tr(\alpha)$

Proof See appendix B.

Corollary 29 If $\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \beta$ is provable then so are $K_i^{\exists}\alpha_1 \wedge \ldots \wedge K_i^{\exists}\alpha_n \rightarrow K_i^{\exists}\beta$ and $K_i^0\alpha_1 \wedge \ldots \wedge K_i^0\alpha_n \rightarrow K_i^{\exists}\beta$.

5.2.3 Knowledge and complexity

We have introduced the concept of algorithmic knowledge in order to represent not only *what* agents know or can know, but also *how long* they need to know what they can know. Our analyses up to now can only answer the first question. By means of the systems presented so far one can infer formulae of the form $K_i^{\exists}\alpha$, but no definite statements of the form $K_i^n\alpha$ or $\neg K_i^n\alpha$. However, to decide if an agent can solve a problem under certain constraints, it is necessary to compute the exact amount of time he needs to solve that problem.

To answer the the "How long"-question, a complexity analysis is needed. The underlying idea is simple. The complexities of many reasoning problem classes are well-known and can be computed at a low cost. (Complexities are typically, but not necessarily, measured by functions of the input size.) Since the average computation speed of an agent can be assumed to be constant, the amount of time he needs to solve some problem can be computed on the basis of the complexity function for the problem's class[4].

Suppose that any formula α in a class $\mathcal{C}$ can be computed at the cost $f(\|\alpha\|)$, where $\|\alpha\|$ is the length of α. Let c_i be a constant number that measures the computation speed of an agent i. If $\alpha \in \mathcal{C}$ and if agent i is able to infer α, then it can be inferred that i is able to compute α within $c_i * f(\|\alpha\|)$ time units. That is, $K_i^{c_i * f(\|\alpha\|)}\alpha$ can be assumed to be true. So, by the aid of complexity theory we can obtain epistemic principles for specific problem classes.

We shall not make any assumption about the nature of the complexity measures and develop our logics independent of the complexity theory in use. We calculate the cost of computing a formula of the language $\mathcal{L}_N^{AK}$ by way of cost functions, which are agent-dependent, partial funtions from $\mathcal{L}_N^{AK}$ to the set of natural numbers. That is, a cost function f_i is defined for each agent i. Intuitively, $f_i(\alpha)$ is the number of time units needed by i to decide α on the basis of all what he knows. Such a cost funtion is defined on the basis of known complexity functions for specific problem classes which can be expressed in the language. Clearly, f_i is defined only for certain sets of formulae, namely for those formulae whose complexities are known. The agents' computation speed and

[4]The computation speed of an agent depends on several factors, e.g., the number of inferences he can perform per time unit, the quality of his algorithms, his ability to classify problems and to select suitable algorithms to solve certain problems, etc. One a time frame has been fixed (i.e., when time units are defined,) the speed can be determined empirically.

the details how the complexity of a certain formula is measured are encapsulated in the specification of the cost function.

If a cost function f_i is defined for a formula, then certain epistemic statements concerning that formula can be made. If the formula α can be inferred reliably by the agent i, then the amount of time needed to infer it is $f_i(\alpha)$, so $K_i^{f_i(\alpha)}\alpha$ can be assumed to be true. Whether or not α can be inferred reliably by i, the introspective knowledge of that can be established after $f_i(\alpha)+1$ time units, because his computation of α will return a positive or negative answer after at most $f_i(\alpha)$ time units. Therefore, $K_i^{\exists}\alpha \rightarrow K_i^{f_i(\alpha)+1}\alpha$ and $\neg K_i^{\exists}\alpha \rightarrow K_i^{f_i(\alpha)+1}\neg K_i^{\exists}\alpha$ are valid. Those axioms will be added to the systems of algorithmic knowledge examined earlier to make more powerful logics for resource-bounded reasoning.

Definition 30 For each $i \in \mathit{Agent}$ let $f_i : \mathcal{L}_N^{AK} \mapsto \omega$ be a cost function for agent i on the language of algorithmic knowledge. A system for resource-bounded reasoning about knowledge is obtained by adding to the logic $\mathbf{K}_N^A$ (or one of its extensions discussed previously) the following axiom schemata:

(AC1) $K_i^{\exists}\alpha \rightarrow K_i^{f_i(\alpha)}\alpha$

(AC2) $K_i^{\exists}\alpha \rightarrow K_i^{f_i(\alpha)+1}K_i^{f_i(\alpha)}\alpha$

(AC3) $\neg K_i^{\exists}\alpha \rightarrow K_i^{f_i(\alpha)+1}\neg K_i^{\exists}\alpha$

provided that $f_i(\alpha)$ is defined.

The complexity analysis makes it possible to prove many unconditional, definite epistemic sentences of the form $K_i^n\alpha$. Let α be a propositional tautology and let $f_i(\alpha)$ be defined. Applying the rule **(NEC^A)** yields the theorem $K_i^{\exists}\alpha$. Hence, $K_i^{f_i(\alpha)}\alpha$ can be derived, by **(AC1)**.

The main task in specifying a system of algorithmic knowledge with complexity analysis is to define the cost functions for the language $\mathcal{L}_N^{AK}$. Let us consider how such cost functions can be constructed. We have argued in section 5.2.1 that if α is provable then $K_i^{\exists}\alpha$ can be inferred. We have posed the question if natural number n can be determined such that the stronger sentence $K_i^n\alpha$ can be assumed as a postulate. It is not yet known whether or not the provability problem for $\mathbf{K}_N^A$ is decidable, so we restrict our attention to certain subclasses. Let α be an objective formula. Then it can be decided in time $2^{\|\alpha\|}$, as we know from complexity theory. If an agent is able to recognize objective formulae and to select a special procedure to compute them, he can derive reliably each objective tautology α in a time proportional to $2^{\|\alpha\|}$.

Interestingly, the previous analysis can be used by an agent within the system in order to reason about himself or about other agents, provided that he has a built-in mechanism to calculate the complexity of reasoning problems. (Such a mechanism can be adopted easily by an agent.) Assume that an agent k tries to find out how long agent i will need to infer a formula α if he chooses to. It is plausible to assume that k can recognize relatively easily that α belongs to the class of objective formulae, so he can reason about agent i exactly like we did before to find out that the time agent i needs is proportional to $2^{\|\alpha\|}$. Generally, k does not know i's computation speed.

However, he may obtain that information from external sources or through his own observations. If k knows the computation speed of i, he will be able to compute the amount of time needed by i to infer α. In other words, he can calculate $f_i(\alpha)$ for any objective formula α. But to estimate the time i would need to derive α, agent k does not have to actually derive it. He has only to recognize the class that α belongs to and then to calculate the complexity of α, which can be accomplished in linear time. So $K_k^{c_k*(\|\alpha\|)}(K_i^{\exists}\alpha \rightarrow K_i^{f_i(\alpha)}\alpha)$ is a plausible postulate, where c_k is the computation speed of k.

Hence, the definition of the complexity function for $i \in \mathit{Agent}$ may include the following clauses: for all objective formulae α

- $f_i(\alpha) = c_i * 2^{\|\alpha\|}$
- $f_i(K_k^n\alpha) = f_i(K_k^{\exists}\alpha) = c_i * \|\alpha\|$ for all $n \in \omega$ and all agents k

The complexity of the decision procedures for normal modal logic has been investigated extensively ([HM92].) It has been shown (cf. theorem 28) that modal epistemic logic can be embedded faithfully to our systems of algorithmic knowledge, i.e., there is a fragment of the language $\mathcal{L}_N^{AK}$ which can be translated one-to-one to modal logic. Consequently, the complexity results obtained for normal modal logic can be applied to determine the cost of solving problems which can be formalized in that fragment. In that way the cost function f_i can be specified for that fragment.

5.2.4 Complexity and the lack of knowledge

It is often important to know not only what an agent knows, but also what he does not know within a certain time limit. An agent's lack of knowledge may restrict his choices, so his action could be predicted or explained accordingly. For instance, consider a rational, utility-maximizing agent which must complete a task under certain time constraint. Moreover, suppose that computing a plan for doing that job is relatively easy, but computing the optimal plan is known to be a very hard problem which cannot be accomplished under the given constraint. (Many optimization problems belong to that category.) Then it is rational to find and execute another plan, and not to attempt to compute the optimal plan at all.

Another example may illustrate our considerations. When we use public-key cryptography to encrypt a message, we want to be sure that someone without the secret key will not be able to know its content within reasonable time — although he can in principle infer it from the public key. The expectation that our message cannot be quickly decrypted is based on the complexity of the reasoning required.

The absence of certain information can be deduced from the presence of other information, utilizing some assumptions about agents' consistency. There is, however, another method for reasoning about the lack of algorithmic knowledge. The expectation that something cannot be known within some time limit is based on the complexity of the reasoning required. We use lower complexity bounds to estimate the least amount of time that an agent would need to infer some sentence, and so to infer what he cannot reliably know within some given limit of time.

For reasoning about what agents cannot infer under some constraints we define for each agent a partial function $f_i' : \mathcal{L}_N^{AK} \mapsto \omega$. Intuitively, $f_i'(\alpha)$ is the minimal amount of time that i needs to compute α. Once such a function has been specified, any formula $\neg K_i^{f_i'(\alpha)}\alpha$ can be assumed, independent of the truth value of $K_i^{\exists}\alpha$. Even if i will eventually succeed to infer α, the computation lasts longer than $f_i'(\alpha)$.

With the ability to reason about algorithmic knowledge or the lack thereof, agents can develop intelligent inference strategies to solve problems under time constraints. The logics of algorithmic knowledge can be implemented and executed directly. When an agent i has to solve a problem α, he checks first if α belongs to a known problem class $\mathcal{C}$. If not, a "universal problem solver" (for any problem that can be described in the language) is activated, and i can only hope to find the solution quickly. But if $\alpha \in \mathcal{C}$, i may estimate its complexity and then decide if the optimal solution can be obtained in time or if some heuristics is needed. Based on that information he can then choose a procedure to solve α. Other agents can also reason about i and about the problems he has to solve to explain or predict his actions accordingly.

5.3 Modeling resources other than time

Until now I have focused solely on a single type of resource, namely time. However, an agent normally needs other resources besides time for solving a problem. For formalizing temporal constraints we have used natural numbers with the standard ordering relation to measure and to compare quantities of the resource time. We have established logical relations between statements built up from formulae of the form $K_i^n\alpha$ ("n time units are sufficient for agent i to compute α".) Now I shall outline how other resources needed for modeling a certain domain can be represented, provided that they are measurable.

Consider situations where m different types of resources are significant, where m is a fixed natural number. We extend the framework of algorithmic knowledge in a natural way. Assume that the resources of each type can be measured using natural numbers (and hence can be compared by means of the standard ordering.) Instead of the one-dimensional time line used previously we consider an m-dimensional resource-space for representing resources. This means that the totality of resources that an agent has at his disposal is represented by an m-tuple $(n_1, \ldots, n_m)$ of m natural numbers. The fact that n_1 unit(s) of resource R_1, n_2 unit(s) of resource R_2, and so on, are sufficient for an agent i to reliably compute α is formalized by the formula $K_i^{n_1,\ldots,n_m}\alpha$. That is, if agent i chooses to compute α and if he has at his disposal n_k unit(s) of resource R_k, for $k = 1, \ldots, m$, then he will succeed in establishing α, consuming no more than the specified amounts of resources. Similarly, the formula $K_i^{\exists}\alpha$ now reads: "agent i is able to compute reliably α using finite amounts of resources."

A meaningful ordering relation on our m-dimensional space can be defined componentwise as follows: $(n_1, \ldots, n_m) \leq (n_1', \ldots, n_m')$ if and only if $n_1 \leq n_1', \ldots, n_m \leq n_m'$. (It can be easily verified that $\leq$ is in fact an ordering relation.) The strict ordering $<$ is defined in the obvious way. It is well-known that $\leq$ and $<$ directed, but not linear. The arguments used in section 5.2 to justify statements about the resource time can be used again to show that similar axioms hold in the case of m resources.

Conclusion

6.1 Summary

One of the principal goals of agent theories is to describe realistic, implementable agents, that is, those which have actually been constructed or are at least in principle implementable. That goal cannot be reached if the inherent resource-boundedness of agents is not treated correctly. Since the modal approach to epistemic logic is not suited to formalize resource-bounded reasoning, the issue of resource-boundedness remains one of the main foundational problems of any agent theory that is developed on the basis of modal epistemic logic.

My work is an attempt to provide theories of agency with a more adequate epistemic foundation. It aims at developing theories of mental concepts that make much more realistic assumptions about agents than other theories. The guiding principle of my theory is that the capacities attributed to agents must be empirically verifiable, that is, it must be possible to construct artificial agents which satisfy the specifications determined by the theory. As a consequence, the unrealistic assumption that agents have unlimited reasoning capacities must be rejected.

In my opinion, resource-bounded reasoning cannot be formalized correctly by restricting the agents' rationality. That is, all attempts to model realistic agents by denying them the use of certain logical rules must be regarded unsatisfactory. The lack of resources does not circumvent an agent from using any of his inference rules. What can be restricted is not the number of logical laws but the number of times they can be applied. Therefore, the correct way to formalize resource-boundedness is to model how the availability of resources (or the lack thereof) can influence an agent's computation.

To achieve the goal of describing resource-bounded agents accurately, the cost of reasoning must be taken seriously. In the thesis I have developed a framework for modeling the relationship between knowledge, reasoning, and the availability of resources. I have argued that the correct form of an axiom for epistemic logic should be: if an agent knows all premises of a valid inference rule and if he performs the right reasoning, then he will know the conclusion as well. Because reasoning requires resources, it cannot be safely assumed that the agent can compute his knowledge if he does not have enough resources to perform the required reasoning. I have demonstrated that on the basis of that idea, the problems of traditional approaches can be avoided and rich epistemic logics can be developed which can account adequately for our intuitions about knowledge.

As a first step, in chapter 4 I have investigated how the explicit concept of knowledge can be represented. I have developed systems of explicit knowledge that can solve the logical omniscience problem of epistemic logic and at the same time account for the agents' full rationality. The agents are non-omniscient, because their actual (or explicit) knowledge at a single time point needs not be closed under any logical law. It is even possible that at some information states they do not know any logical truth at all. On the other hand, they are non-ignorant, because they are capable of logical thinking. They can use their reasoning capacities to infer new information from what they already know. Their rationality is not restricted by any artificial, ad hoc postulate saying that their inference mechanisms are incomplete.

In the next step (chapter 5) I have introduced algorithmic knowledge — a concept of knowledge that is suited for establishing direct relations between and agent's available resources and his knowledge. I have argued that the proposed algorithmic concept of knowledge can serve as a basis for action. The main idea is to consider how much resources an agent will need to compute the answer to a certain query. That question can be answered by combining epistemic logic with a complexity analysis. Following this strategy I have developed systems for reasoning about algorithmic knowledge which can describe non-omniscient, albeit fully rational agents. Moreover, the defined systems have enough expressive power to formalize quantitative constraints.

6.2 Related works

My work is complementary to other approaches to resource-bounded agents (e.g., [Kor98]) in the following sense: instead of trying to find near-optimal solutions to some specific problem (or class of problems) I try to model the control mechanism used by an agent to select a suitable action sequence for the given situation. Such a mechanism could be used, e.g., to decide if in a certain situation it is necessary or desirable to trade quality of results for time or other resources.

Most theories of agency have tried to integrate knowledge and time in a single framework. However, in most cases some modal epistemic logic is combined with some temporal logic, yielding a hybrid system that can at best be used to describe implicit knowledge. There are in fact very few works that treat time as a valuable resource for computing knowledge. In the following I shall discuss briefly some other attempts to investigate the relation between knowledge and reasoning actions and the evolution of knowledge over time.

Elgot-Drapkin et. al. ([EDMP91], [NKP94]) developed what they called step-logics (later renamed as active-logics) to model the reasoning of agents over time. The underlying intuition of their approach is that an agent can carry out one step of reasoning at each time step: if two premises of a rule are known at some point then the agent will apply the rule to know the conclusion at the next point. For example, if both α and β are known at time t then their conjunction is known at time $t+1$, so the formula $K_i^t\alpha \wedge K_i^t\beta \rightarrow K_i^{t+1}(\alpha \wedge \beta)$ is assumed as an axion.

Although the step-logics framework takes the cost of reasoning into account, this is not done consequently. Therefore, the assumptions about the agents' reasoning capacities are too strong. For example, the knowledge necessitation rule ("agents believe

all logical truths") and the congruence rule ("agents believe all logically equivalent formulae of his beliefs") are valid. Moreover, if $\alpha \to \beta$ is provable and α is known at time t, then β is known at time $t+1$, however complex the derivation of β from α may be. Finally, unlimited space and parallelism must be assumed implicitly in order to justify an axiom like $K_i^t\alpha \wedge K_i^t\beta \to K_i^{t+1}(\alpha \wedge \beta)$: it is supposed that any logical consequence which can be derived in one step is actually derived after one time unit. Thus, the step-logics framework suffers from several strong forms of logical omniscience.

Stelzner ([Ste84]) discussed a number of epistemic concepts, their role in rational discourses and their dependency on parameters such as time, agent, context. He proposed a family of logics to formalize the concept of a (hypothetical) obligation to defend some asserted sentence. That concept is related to the concepts of knowledge and belief in the following way. In a rational discourse, if an agent asserts some sentence, then he has the obligation to defend it when it is challenged, because he has made public through his assertion that he believes in the sentence. The obligation to defend the sentence is only hypothetical, because it does not arise if the agent's assertion is not challenged.

To qualify as rational, the agents in a discourse must satisfy certain conditions. A rationality postulate may say, for example, that if an agent is obligated to defend α at time t and if β can be inferred from α by one inference step, then the agent can be obligated to defend β at time $t+1$. Hence, $K_i^t(\alpha \wedge \beta) \to K_i^{t+1}\alpha$ is assumed as an axiom. (A time line isomorphic to the set of natural numbers, generated by the consecutive "moves" in the discourse, is assumed.) With the aid of such axioms one can classify agents according to their rationality.

If interpreted as logics of knowledge, Stelzner's logics could be regarded as formalizations of the concept of implicit knowledge, but not of explicit knowledge. A statement such as $K_i^t(\alpha \wedge \beta) \to K_i^{t+1}\alpha$ may be more acceptable than the axiom $K_i^t(\alpha \wedge \beta) \to K_i^t(\alpha)$, but it is still too strong for the notion of actual knowledge.

Halpern, Moses, and Vardi ([HMV94]) developed a general framework for modeling knowledge and computation. It is assumed that at any state, an agent has some local data and exactly one local algorithm which is defined for all formulae and always terminates with one of three possible answers "Yes", "No", or "?". Intuitively, "Yes" means that the formula in question is the agent's implicit knowledge, "No" means that it is not, and the answer "?" means that the agent is unable to determine whether or not the formula follows from all what he knows. At any state, the agent is said to know a fact if the output of his local algorithm is "Yes" when running on that fact and his local data. In other words, he can compute that he (implicitly) knows that fact. This notion of knowledge is called algorithmic knowledge by the authors. A local algorithm of an agent i is said to be sound if for any formula α and any local data, the answer "Yes" implies that α is in fact i's implicit knowledge at the state in consideration, and the answer "No" implies that he does not know α implicitly at that state. The algorithm is called complete if it does not return the answer "?".

Obviously, if the employed algorithms are not complete then logical omniscience is avoided, so some aspect of resource boundedness can be modeled. The authors view their concept of knowledge as a kind of explicit knowledge. However, an agent cannot really act upon that knowledge immediately because that information must be inferred first. Hence, that kind of knowledge may not suffice to justify an agent's actions if he

needs to act before the computation is completed. Moreover, under certain circumstances that concept of knowledge exhibits certain properties of implicit knowledge. In fact, as the authors pointed out, their notion of algorithmic knowledge coincides with implicit knowledge when sound and complete algorithms are employed.

The framework of Halpern et. al. specifies a general epistemic language which can be used to describe a large number of situations where computing knowledge is involved. However, it does not really specify a logic for reasoning about knowledge: because their notion of an algorithm is too general, their class of models is too large, therefore no genuine epistemic statement is valid in all models. There seems to be no easy way to make their concept of knowledge more specific so that epistemic inference relations can be established. My framework differs from that of [HMV94] in that aspect: I have shown that certain epistemic statements are valid for intelligent, resource-bounded agents. The epistemic consequence relations defined in my framework justify inferences about agents once a general rationality assumption has been made.

In the literature on belief revision some authors have considered belief-changing actions. For example, Van Linder, van der Hoek and Meyer ([vLvdHM95a], [vLvdHM95b]) have done some work to formalize the change of knowledge through actions. However, they made very strong assumptions about knowledge: their agents are logically omniscient. The actions they consider lead from one deductively closed belief set to another. Thus, their work should be read in terms of information dynamics, and not knowledge dynamics.

6.3 Future directions

The proof systems defined for explicit knowledge (chapter 4) and algorithmic knowledge (chapter 5) provide a procedural semantics for these concepts. It remains an open problem to develop an intuitively acceptable declarative semantics for them. Although it is possible to develop a model theory along the lines of [Ho95] and to prove completeness of the specified systems with respect to those models, such a model theory is simply a reformulation of the procedural semantics and not very useful. It does not provide us with a tool to determine if all (semantically) valid epistemic statements have been captured by the proof system, and it does not allow us to analyze the concepts of knowledge and belief in terms of simpler, more fundamental concepts. But this is exactly what a well-motivated and intuitively acceptable semantics should do.

Semantics has also proved helpful for studying the complexity of modal epistemic logic. The complexity analysis for the specified logics of explicit and algorithmic knowledge remains an open issue. It is hoped that their complexities can be determined by embedding them into systems whose complexities are known.

The investigated logics of explicit knowledge ($\mathbf{DEK}_N$ and its extensions) have been monotonic in two aspects. First, their consequence operations are monotonic. Second, the knowledge of the agents always grows over time. A very interesting, still open problem is to develop logics of explicit knowledge based on non-monotonic logic, where the agents can revise their knowledge when they find out that their knowledge is inconsistent. We may expect to find interesting connections with two other, very active fields of AI research, viz. to non-monotonic reasoning and to the logic of belief revision.

This seems to be a promising field of research and needs further investigations.

My study of algorithmic knowledge has concentrated on a single time point in the real time line. A formula like $K_i^n\alpha$ is primarily a statement about the agent's i current ability to compute knowledge. It does not say anything about his actual knowledge n time units from now — except for the case $n = 0$. An interesting problem is to relate an agent's algorithmic knowledge at different time points to each other. Let me elaborate that issue in some more details.

Let the integers be the representation of the objective time structure ("real time line", "real world history"). For any integer t we shall write $K_{i,t}^n\alpha$ instead of $K_i^n\alpha$ to make clear that t is the time point under consideration. That is, if at time t the agent i starts computing α then he will needs at most n time units to complete the task, so he will succeed at $t+n$ at the latest. The question is how the inference relations between an agent's algorithmic knowledge at two different points can be formalized.

Consider the simple case $n = 0$ first. Can we assume that an agent's explicit knowledge persists over time? For example, is the formula $K_{i,t}^0\alpha \rightarrow K_{i,t+1}^0\alpha$ valid? As I have argued in chapter 4, such a persistence axiom cannot be assumed for all formulae. So we may ask under which circumstances those persistence axioms are valid, or if certain default inference rules can be used for reasoning about the evolution of explicit knowledge over time.

A related question is how an agent's reasoning capacities change over time. If at time t he needs n time units to compute α, will it remain true at time $t+1$? That is, is the formula $K_{i,t}^n\alpha \rightarrow K_{i,t+1}^n\alpha$ valid? This is apparently not true universally, since the truth-value of α may not be the same at different states. Moreover, at time $t+1$ the agent i may need more time to compute the same query. Nevertheless, it seems plausible to assume that an agent's reasoning capacities do not decrease over time. Under this assumption, the mentioned formula may be valid if α has a certain syntactic structure, e.g., if α is objective and does not contain any negation sign.

Finally, the connection between $K_{i,t}^n\alpha$ and $K_{i,t+n}^0\alpha$ is worth examining. The former formula says that α is implicit knowledge of agent i at time t, and he has the capacity to make it explicit if he chooses to do so and if he has enough resources to carry out his computation, in this case n time units. The latter formula says that α has actually become explicit knowledge of i at time $t+n$. A framework for reasoning about knowledge and action should be able to formalize the agent's i computation of α between t and $t+n$.

To summarize, an agent's algorithmic knowledge at different moments can be related to each other through certain consequence relations. The concrete conditions under which such inferences may be drawn must be examined carefully. It can be expected that default logic and other mechanisms of non-monotonic reasoning can help to specify consequence relations for algorithmic knowledge.

Propositional modal, temporal, and dynamic logic

A.1 Modal logic

In this section we shall review modal logic briefly. We shall only define the syntax and semantics of the basic propositional systems and state without proofs some of the most basic results about them. More complete overviews of the subject can be found in [HC96], [Che80], or [Gol87].

A.1.1 Sytax of modal logic

The language of propositional modal logic is formed by extending that of the propositional calculus by an one-place modal operator $\Box$. Formally:

Definition 31 (Modal language) Let *Atom* be a nonempty set of atomic formulae. $\mathcal{L}^M$ is the least set such that

1. $Atom \subseteq \mathcal{L}^M$
2. If $\alpha \in \mathcal{L}^M$ then $\neg\alpha \in \mathcal{L}^M$
3. If $\alpha \in \mathcal{L}^M$ and $\beta \in \mathcal{L}^M$ then $(\alpha \to \beta) \in \mathcal{L}^M$
4. If $\alpha \in \mathcal{L}^M$ then $\Box\alpha \in \mathcal{L}^M$

The formula $\Box\alpha$ is read: it is necessarily true that α. The possibility operator is introduced as an abbreviation: $\Diamond\alpha =_{def} \neg\Box\neg\alpha$. If $\Gamma \subseteq \mathcal{L}^M$ is a set of formulae then $\Box(\Gamma)$ denotes the set $\{\Box\alpha : \alpha \in \Gamma\}$. We stipulate that the modal operators bind more strongly than the Boolean connectives. Furthermore, we introduce the following abbreviations:

$$\Box^0\alpha =_{def} \alpha$$

$$\Box^{m+1}\alpha =_{def} \Box(\Box^m\alpha)$$

for all $m \geq 0$ and $\alpha \in \mathcal{L}^M$.

All the modal systems we consider are formed by adding to a complete axiomatization of the propositional calculus some specific modal axiom schemata and rules of inference. We shall consider some modal logics determined by axioms and rules from the following lists:

(PC) All propositional tautologies

(MP) Modus ponens: from α and $\alpha \to \beta$ to infer β

(K) $\Box\alpha \wedge \Box(\alpha \to \beta) \to \Box\beta$

(T) $\Box\alpha \to \alpha$

(D) $\Box\alpha \to \neg\Box\neg\alpha$

(4) $\Box\alpha \to \Box\Box\alpha$

(5) $\neg\Box\alpha \to \Box\neg\Box\alpha$

(N) $\Box(\alpha \vee \neg\alpha)$

(C) $\Box\alpha \wedge \Box\beta \to \Box(\alpha \wedge \beta)$

(G) $\Diamond\Box\alpha \to \Box\Diamond\alpha$

(NEC) From α to infer $\Box\alpha$ (Necessitation)

(MON) From $\alpha \to \beta$ to infer $\Box\alpha \to \Box\beta$ (Monotony)

(CGR) From $\alpha \leftrightarrow \beta$ to infer $\Box\alpha \leftrightarrow \Box\beta$ (Congruence)

(RK$_n$) From $\alpha_1 \wedge \ldots \wedge \alpha_n \to \beta$ to infer $\Box\alpha_1 \wedge \ldots \wedge \Box\alpha_n \to \Box\beta$, for all $n \in \omega$

Instead of using **(PC)** for defining a logic to include all tautologies, it would suffice to include a set of schemata from which all tautologies can be derived by appropriate rules of inference, e.g., modus ponens. An example of such a set of schemata is:

(PC1) $\alpha \to (\beta \to \alpha)$

(PC2) $(\alpha \to (\beta \to \gamma)) \to ((\alpha \to \beta) \to (\alpha \to \gamma))$

(PC3) $(\neg\beta \to \neg\alpha) \to (\alpha \to \beta)$

Let **(A)** be one of the axiom schemata listed above and α the formula named **(A)**. Then $\Box^\omega$**(A)** denotes the set $\{\Box^i\alpha : i \in \omega\}$. For example, $\Box^\omega$**(K)** is the set $\{\Box^i(\Box\alpha \wedge \Box(\alpha \to \beta) \to \Box\beta) : i \in \omega\}$.

A modal logic (over the language $\mathcal{L}^M$) is called classical if it is closed under the rule of congruence **(CGR)**. The minimal classical logic, which is axiomatized by **(PC)**, **(MP)**, and **(CGR)**, is denoted **E**. A modal system is called monotonic if it is closed under the monotony rule **(MON)**. The minimal monotonic logic, which is axiomatized by **(PC)**, **(MP)**, and **(MON)**, is called **M**. A modal logic is called normal if it contains the schema **(K)** and is closed under the rule of necessitation **(NEC)**. Some equivalent axiomatizations of the minimal normal modal logic **K** are given in the following.

Lemma 32 The following axiom systems are equivalent axiomatizations of the logic **K**:

- **(PC)**, **(MP)**, **(K)**, and **(NEC)**.
- **(PC)**, **(MP)**, and **(RK$_n$)**
- **(PC)**, **(MP)**, **(C)**, **(N)**, and **(MON)**
- **(PC)**, **(MP)**, **(K)**, **(N)**, and **(CGR)**
- $\Box^\omega$**(PC1)**, $\Box^\omega$**(PC2)**, $\Box^\omega$**(PC3)**, $\Box^\omega$**(K)**, and **(MP)**

The last of the above axiomatizations is less common than the other. It is however of interest for epistemic logic because none of the inference rules **(NEC)**, **(MON)**, and **(CGR)** is required.

It is easy to see that every monotonic logic is also classical, and every normal logic is also monotonic logic and classical. In particular, one can show that **E** is a proper subsystem of **M**, and **M** is in turn a proper subsystem of **K**, i.e., **E** $\subset$ **M** $\subset$ **K**.

Additional normal (monotonic, classical) systems of modal logic can be formed by adding axioms to the basic systems **K** (**M**, **E**). The new systems are often named by appending the names of the additional modal axioms used to the name of the basic system, e.g., **EK** is the logic **E** together with the axiom **K**; the logic **EK** together with the axiom **4** is called **EK4**; and **KD45** is the logic **K** together with the axioms **(D)**, **(4)**, and **(5)**. Some modal logics are better known under their historical names, in particular, **KT4** is known as **S4**, **KT4G** as **S4.2**, and **KT45** as **S5**.

A.1.2 Semantics for normal modal logic: Kripke models

Normal modal logics can be given a nice semantics by means of Kripke models, also known as possible worlds semantics.

Definition 33 (Semantics for normal modal logics) A Kripke model is a structure $M = (S, R, V)$ where

1. S is a nonempty set (called the set of possible worlds, or states),
2. $R \subseteq S \times S$ is a binary relation on S (called the accessibility relation, or alternativeness relation), and
3. V is a valuation function $V : Atom \mapsto Pow(S)$.

The satisfaction relation $\models$ is defined recursively over the complexity of formulae as follows:

- $M, s \models \phi$ iff $s \in V(\phi)$, for all atomic formulae $\phi \in Atom$
- $M, s \models \neg\alpha$ iff $M, s \not\models \alpha$, i.e., it is not the case that $M, s \models \alpha$
- $M, s \models \alpha \rightarrow \beta$ iff $M, s \not\models \alpha$ or $M, s \models \beta$

- $M, s \models \Box\alpha$ iff for all $t \in S$, sRt implies $M, t \models \alpha$

$M, s \models \alpha$ is read: α is true (or satisfied) at state s in model M.

Let $\mathcal{C}$ be a class of models for the language $\mathcal{L}^M$. A formula $\alpha \in \mathcal{L}^M$ is said to be satisfiable with respect to $\mathcal{C}$ if it is true at some state in some model in $\mathcal{C}$; otherwise it is unsatisfiable. The reference to the class $\mathcal{C}$ of models can be omitted if it is implicitly understood which class is meant. A formula α is called valid in a model M, in symbol $M \models \alpha$, if it is true at all states of M. It is called valid with respect to some class of models $\mathcal{C}$, denoted $\models_{\mathcal{C}} \alpha$, if it is valid in all models of that class.

Probably the most important reason for the popularity of possible-worlds semantics is that common modal axioms correspond exactly to certain algebraic properties of Kripke models in the following sense: an axiom is valid in a model M if and only if the alternativeness relation of M satisfies some algebraic condition. (In fact, the correspondence holds on a higher abstraction level, the level of *frames*, consult [vB84] for details.) In particular:

- **(T)** holds iff R is reflexive
- **(D)** holds iff R is serial
- **(4)** holds iff R is transitive
- **(5)** holds iff R is Euclidean
- **(G)** holds iff R is directed

The common normal modal logics can be characterized by appropriate classes of Kripke models. In the following theorem, a Kripke model is said to be reflexive iff its accessibility relation is reflexive, and so on.

Theorem 34 1. The minimal normal system **K** is determined by the class of all Kripke models: $\vdash_{\mathbf{K}} \alpha$ iff α is valid in all Kripke models.

2. **KT** is determined by the class of reflexive Kripke models
3. **KD** is determined by the class of serial Kripke models
4. **S4** is sound and complete wrt the class of reflexive, transitive models.
5. **S5** is sound and complete wrt the class of models where the accessibility relation is an eqivalence relation.
6. **KD4** is determined by the class of serial and transitive models.
7. **KD45** is sound and complete wrt the class of serial, transitive, and Euclidean models.

The common normal propositional modal logics are conservative extensions of classical logic: if a formula α does not contain any occurrence of the modal operator then it is provable in a system mentioned in the previous theorem if and only if it is provable in the propositional calculus.

A.1.3 Montague-Scott semantics

Kripke models as defined above cannot account for non-normal modal logics. To develop an adequate semantics for classical (and monotonic) modal logics we need a generalization of Kripke sematics, the so-called neighborhood semantics (also known as Montague semantics, or Montague-Scott semantics). A complete overview of the basic model theory of classical systems is found in [Che80].

Definition 35 (Semantics for classical systems) A neighborhood model is a structure $M = (S, N, V)$ where

1. S is a nonempty set, called the set of worlds
2. $N : S \mapsto Pow(Pow(S))$ is a function from S to the powerset of the powerset of S.
3. $V : Atom \mapsto S$ is a valuation function

Satisfaction is defined as in definition 33 except that:

- $M, s \models \Box\alpha$ iff $\{t \in S : M, t \models \alpha\} \in N(s)$

Intuitively, $N(s)$ consists of the intensions of all formulae which are necessary at s, where the intension of a formulae is the set of all worlds where it is true. Thus, something is necessarily true at a world if and only if its intension is contained in the set of intensions of formulae considered necessary at that world.

Theorem 36 The minimal classical system **E** is sound and complete wrt the class of all neighborhood models.

Semantics for extensions of **E**, including the common monotonic and normal logics, can be obtained by restricting the class of neighborhood model through appropriate conditions ([Che80]). For example, **EK** is determined by the class of all neighborhood models satisfying the condition: for all $X, Y \subseteq S$ and $s \in S$, if $(S \setminus X) \cup Y \in N(s)$ and $X \in N(s)$ then $Y \in N(s)$.

A.1.4 Basic temporal logic

The reading of the operator $\Box$ as "always in the future" (and accordingly, $\Diamond$ as "sometimes in the future") has proved plausible for many modal logics. Those systems are the most simple temporal (or tense) logics. Semantically, the "possible worlds" of a Kripke model are interpreted as moments in time, and the accessibility relation is viewed as the the relation "later than".

Clearly, not all algebraic properties that can be imposed on binary relations are meaningful under a temporal interpretation. The most interesting properties are those characterizing ordering relations. Transitivity is probably the most basic condition that can be placed on the relation "later than", so it is typically assumed that temporal structures are transitive. Accordingly, the minimal temporal logic is axiomatized by

adding to system **K** the axiom **(4)**, which corresponds to transitivity. That system is denoted $\mathbf{K}_t\mathbf{4}$, with the subscript t indicating that a temporal logic is being considered.

More complex logics of time are usually developed on the basis of richer languages. Typically, some past operators are also considered. Different, often incompatible models of time can be developed by adopting appropriate axioms. For example, time can be assumed to be linear or branching, discrete or dense, limited or unlimited. For more complete overviews consult [Bur84], [Eme90].

A.2 Propositional Dynamic Logic

The formal language of Propositional Dynamic Logic (PDL) is built up from two sets of primitive symbols, a countable set At of atomic formulae and a countable set Prg of atomic programs. The set of formulae is the least set containing At and is closed under the rules: if α and β are formulae then $\neg\alpha$ and $\alpha \rightarrow \beta$ are formulae; if α is a formula and x is a program then $[x]\alpha$ is a formula. The set of programs is the least set containing all atomic programs and closed under the rule: if x and y are programs then $x;y$, $x \cup y$ and x^* are programs; if α is a formula then $\alpha?$ is a program.

The dual operator to $[x]$ is defined as follows:

$$\langle x\rangle\alpha =_{def} \neg[x]\neg\alpha$$

The intuitive reading of the program connectives and the formulae are as follows:

- $x;y$: "perform x and then y" (composition)
- $x \cup y$: "do either x or y" (non-deterministic choice)
- x^*: "perform x finitely many (including zero) times" (iteration)
- $\alpha?$: "test if α currently holds"
- $[x]\alpha$: "α always holds after x is performed"
- $\langle x\rangle\alpha$: "sometimes after x is performed, α holds".

A PDL-model is a structure $M = (W, R, V)$, where W is a non-empty set, R is a function which assigns to every program x a binary relation $R(x)$ on W, and V is a valuation function which assigns to every atomic formula ϕ a subset $V(\phi)$ of W.

The satisfaction relation "α is true at point s in model M", denoted $M, s \models \alpha$, is defined inductively on the formation of α as follows:

1. $M, s \models \phi$ iff $s \in V(\phi)$ for any atomic formula ϕ
2. $M, s \models \neg\alpha$ iff $M, s \not\models \alpha$, i.e., it is not the case that $M, s \models \alpha$
3. $M, s \models \alpha \rightarrow \beta$ iff $M, s \not\models \alpha$ or $M, s \models \beta$
4. $M, s \models [x]\alpha$ iff for all $t \in W$, if $(s, t) \in R(x)$ then $M, s \models \alpha$

To preserve the intuitive reading of the program connectives, the following formal requirements must be met:

$$R(x;y) = R(x);R(y)$$

$$R(x \cup y) = R(x) \cup R(y)$$

$$R(x^*) = (R(x))^*$$

$$R(\alpha?) = \{(s,s) \in W \times W : M, s \models \alpha\}$$

Well-known results say that PDL is decidable and finitely axiomatizable (cf. [Har84], [Gol87], [KT90]). A complete axiomatization consists of the following axioms and inference rules:

(A0) All propositional tautologies

(A1) $[x]\alpha \wedge [x](\alpha \to \beta) \to [x]\beta$

(A2) $[x;y]\alpha \leftrightarrow [x][y]\alpha$

(A3) $[x \cup y]\alpha \leftrightarrow [x]\alpha \wedge [y]\alpha$

(A4) $[x^*]\alpha \to \alpha \wedge [x]\alpha$

(A5) $[x^*]\alpha \to [x^*][x^*]\alpha$

(A6) $[x^*](\alpha \to [x]\alpha) \to (\alpha \to [x^*]\alpha)$

(A7) $[\alpha?]\beta \leftrightarrow (\alpha \to \beta)$

(R1) If α and $\alpha \to \beta$ are theorems, then β is a theorem

(R2) If α is a theorem, then so is $[x]\alpha$

It can be shown that the axiom **(A5)** is still valid if $R(x^+) = (R(x))^+$, i.e., if the program construct x^* is interpreted by the transitive closure of the relation corresponding to x, and not by the reflexive, transitive one. This fact explains how the minimal temporal logic of transitive time $\mathbf{K}_t\mathbf{4}$ (with only the future operators) can be embedded into dynamic logic: the temporal operator "always in the future" is interpreted by the operator $[x^*]$, where x is any (atomic) program.

Formal Proofs

Theorem 21

1. $K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$

(1)	$K_i\alpha \to [F_i]K_i\alpha$	**(DE2)**
(2)	$K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to [F_i]K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta)$	(1)
(3)	$[F_i]K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$	$\mathbf{K}_t\mathbf{4}$
(4)	$K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$	(2), (3)

2. $K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$

(1)	$K_i\alpha \to [F_i]K_i\alpha$	**(DE2)**
(2)	$K_i\alpha \wedge \langle F_i\rangle K_i\beta \to [F_i]K_i\alpha \wedge \langle F_i\rangle K_i\beta$	(1)
(3)	$[F_i]K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$	$\mathbf{K}_t\mathbf{4}$
(4)	$K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$	(2), (3)

3. $K_i\alpha \wedge K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$

(1)	$K_i\beta \to \langle F_i\rangle K_i\beta$	$\mathbf{(MON}_{de}\mathbf{)}$
(2)	$K_i\alpha \wedge K_i\beta \to K_i\alpha \wedge \langle F_i\rangle K_i\beta$	(1)
(3)	$K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$	Th. 21.2.
(4)	$K_i\alpha \wedge K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$	(2), (3)

4. $K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$

(1)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle K_i\alpha$	$\mathbf{(MON}_{de}\mathbf{)}$
(2)	$K_i(\alpha \wedge \beta) \to [F_i]K_i(\alpha \wedge \beta)$	**(DE2)**
(3)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle K_i\alpha \wedge [F_i]K_i(\alpha \wedge \beta)$	(1), (2)
(4)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i(\alpha \wedge \beta))$	(3), $\mathbf{K}_t\mathbf{4}$
(5)	$K_i\alpha \to [F_i]K_i\alpha$	**(DE2)**
(6)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle K_i\beta$	$\mathbf{(MON}_{de}\mathbf{)}$
(7)	$K_i\alpha \wedge K_i(\alpha \wedge \beta) \to [F_i]K_i\alpha \wedge \langle F_i\rangle K_i\beta$	(5), (6)
(8)	$[F_i]K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	$\mathbf{K}_t\mathbf{4}$
(9)	$K_i\alpha \wedge K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(7), (8)
(10)	$\langle F_i\rangle(K_i\alpha \wedge K_i(\alpha \wedge \beta)) \to \langle F_i\rangle\langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(9), $\mathbf{K}_t\mathbf{4}$
(11)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle\langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(4), (10)
(12)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(11), $\mathbf{K}_t\mathbf{4}$

5. $\langle F_i\rangle K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$

(1)	$K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	Th. 21.4.
(2)	$\langle F_i\rangle K_i(\alpha \wedge \beta) \to \langle F_i\rangle\langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(1), $\mathbf{K}_t\mathbf{4}$
(3)	$\langle F_i\rangle K_i(\alpha \wedge \beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(2), $\mathbf{K}_t\mathbf{4}$

Theorem 22

1. $\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$.

(1)	$\langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle[F_i]K_i(\alpha \to \beta)$	**(DE2)**, $\mathbf{K}_t\mathbf{4}$
(2)	$\langle F_i\rangle K_i(\alpha \to \beta) \to [F_i]\langle F_i\rangle K_i(\alpha \to \beta)$	(1), **(TL3)**
(3)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\alpha \wedge [F_i]\langle F_i\rangle K_i(\alpha \to \beta)$	(2)
(4)	$\langle F_i\rangle K_i\alpha \wedge [F_i]\langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta))$	$\mathbf{K}_t\mathbf{4}$
(5)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta))$	(3), (4)
(6)	$K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$	Th. 21
(7)	$\langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta)) \to \langle F_i\rangle\langle F_i\rangle K_i\beta$	(6), $\mathbf{K}_t\mathbf{4}$
(8)	$\langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta)) \to \langle F_i\rangle\langle F_i\rangle K_i\beta$	(5), (7)
(9)	$\langle F_i\rangle\langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i\beta$	$\mathbf{K}_t\mathbf{4}$
(10)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$	(8), (9)

2. $\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i(\alpha \wedge \beta)$

(1)	$\langle F_i\rangle K_i\beta \to \langle F_i\rangle[F_i]K_i\beta$	**(DE2)**, $\mathbf{K}_t\mathbf{4}$
(2)	$\langle F_i\rangle K_i\beta \to [F_i]\langle F_i\rangle K_i\beta$	(1), **(TL3)**
(3)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle K_i\alpha \wedge [F_i]\langle F_i\rangle K_i\beta$	(2)
(4)	$\langle F_i\rangle K_i\alpha \wedge [F_i]\langle F_i\rangle K_i\beta \to \langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i\beta)$	$\mathbf{K}_t\mathbf{4}$
(5)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i\beta)$	(3), (4)
(6)	$K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	Th. 21
(7)	$\langle F_i\rangle(K_i\alpha \wedge \langle F_i\rangle K_i\beta) \to \langle F_i\rangle\langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(6), $\mathbf{K}_t\mathbf{4}$
(8)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i\beta \to \langle F_i\rangle\langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	(5), (7)
(9)	$\langle F_i\rangle\langle F_i\rangle(K_i\alpha \wedge K_i\beta) \to \langle F_i\rangle(K_i\alpha \wedge K_i\beta)$	$\mathbf{K}_t\mathbf{4}$
(10)	$\langle F_i\rangle K_i\alpha \wedge \langle F_i\rangle K_i(\alpha \to \beta) \to \langle F_i\rangle K_i\beta$	(8), (9)

Theorem 28

We prove the first correspondence. That is, we shall show that α is $\mathbf{K}_N$-provable if and only if $tr(\alpha)$ is $\mathbf{K}_N^A$-provable. The other results are obtained in the same way.

First, we show that the translation of every $\mathbf{K}_N$-theorem is a $\mathbf{K}_N^A$-theorem. Let α be a theorem of $\mathbf{K}_N$. We show inductively over the length of the proof of α in $\mathbf{K}_N$ that $tr(\alpha)$ is a theorem of $\mathbf{K}_N^A$.

Assume that α is one of $\mathbf{K}_N$-axioms, i.e., it is either a propositional tautology or an instance of the schema $\mathbf{K}$. In the former case $tr(\alpha)$ is also a propositional tautology, and in the latter case $tr(\alpha)$ is an instance of $(\mathbf{K}^A)$, so in any case $tr(\alpha)$ is a $\mathbf{K}_N^A$-theorem.

Now suppose that α has been derived by applying modus ponens to β and $\beta \to \alpha$. By induction hypothesis, both $tr(\beta)$ and $tr(\beta \to \alpha)$ are $\mathbf{K}_N^A$-theorems. By definition, $tr(\beta \to \alpha)$ is $tr(\beta) \to tr(\alpha)$, so in $\mathbf{K}_N^A$ we can apply modus ponens to infer $tr(\alpha)$.

Finally, suppose that α has been derived from β using the knowledge necessitation rule **(NEC)**, i.e., α is $K_i\beta$ and β is a $\mathbf{K}_N$-theorem. Then $tr(\beta)$ is a theorem of $\mathbf{K}_N^A$,

by induction hypothesis. According to the definition of tr, $tr(\alpha)$ is $K_i^{\exists} tr(\beta)$, which can be derived from the theorem $tr(\beta)$ by applying the rule **(NEC^A)**. Hence, $tr(\alpha)$ is a $\mathbf{K}_N^A$-theorem.

To prove the converse, we define the inverse function $tr^{-1} : \mathcal{L}_N^{AK} \mapsto \mathcal{L}_N^K$ of tr as follows:

- $tr^{-1}(\phi) = \phi$ for all $\phi \in Atom$
- $tr^{-1}(\neg\alpha) = \neg tr^{-1}(\alpha)$
- $tr^{-1}(\alpha \to \beta) = tr^{-1}(\alpha) \to tr^{-1}(\beta)$
- $tr^{-1}(K_i^n \alpha) = K_i(tr^{-1}(\alpha))$, for all $n \in \omega$
- $tr^{-1}(K_i^{\exists} \alpha) = K_i(tr^{-1}(\alpha))$

It is easy to see that $tr^{-1}(tr(\alpha)) = \alpha$ for all $\alpha \in \mathcal{L}_N^K$. However, the converse does not hold: generally $tr(tr^{-1}(\alpha))$ and α are different formulae. For instance, $tr(tr^{-1}(K_i^n \phi)) = tr(K_i(tr^{-1}(\phi))) = tr(K_i \phi) = K_i^{\exists} \phi \neq K_i^n \phi$.

Assume that $tr(\alpha)$ is provable in $\mathbf{K}_N^A$. Then there is a $\mathbf{K}_N^A$-proof leading to $tr(\alpha)$, i.e., there is a sequence $\beta_0, \ldots, \beta_n$ of $\mathbf{K}_N^A$-formulae such that β_n is $tr(\alpha)$ and each β_k $(k = 0, \ldots, n)$ is either an axiom of $\mathbf{K}_N^A$ or is derived from previous formulae in the sequence using one of the inference rules. We show that the sequence $tr^{-1}(\beta_0), \ldots, tr^{-1}(\beta_n)$ is a proof of α in $\mathbf{K}_N$.

Suppose that β_k is axiom of $\mathbf{K}_N^A$. If β_k is a propositional tautology then so is $tr^{-1}(\beta_k)$. If β_k is an instance of **(K^A)** then $tr^{-1}(\beta_k)$ is an instance of **(K)**. If β_k is an instance of **(P^A)** or **(Q^A)** then $tr^{-1}(\beta_k)$ is a propositional tautology.

Suppose that β_k is derived from β_l and $\beta_l \to \beta_k$ by modus ponens. Then $tr^{-1}(\beta_l)$ and $tr^{-1}(\beta_l \to \beta_k)$ are $\mathbf{K}_N$-theorems, so $tr^{-1}(\beta_k)$ can be inferred from $tr^{-1}(\beta_l)$ and $tr^{-1}(\beta_l) \to tr^{-1}(\beta_k)$ (which is by definition the same formula as $tr^{-1}(\beta_l \to \beta_k)$.)

Finally, let β_k be derived from β_l by applying **(NEC^A)**. Then β_k is $K_i^{\exists} \beta_l$, hence $tr^{-1}(\beta_k)$ is $K(tr^{-1}(\beta_l))$, which can be derived from $tr^{-1}(\beta_l)$ using **(NEC)**.

Thus, in any case $tr^{-1}(\beta_k)$ is a theorem of $\mathbf{K}_N$. As observed earlier, $tr^{-1}(\beta_n) = tr^{-1}(tr(\alpha)) = \alpha$. So α is a theorem of $\mathbf{K}_N$. This completes the proof.

Bibliography

[And58] A. R. Anderson, *A reduction of deontic logic to alethic modal logic*, Mind **67** (1958), 100–103.

[And67] A. R. Anderson, *The formal analysis of normative systems*, The Logic of Decision and Action (N. Rescher, ed.), University of Pittsburgh Press, 1967, pp. 147–213.

[Aum76] R. Aumann, *Agreeing to disagree*, Annals of Statistics **4** (1976), no. 6, 1236–1239.

[Bac94] M. Bacharach, *The epistemic structure of a theory of a game*, Theory and Decision **37** (1994), no. 1, 7–48.

[Bar89] J. Barwise, *On the model theory of common knowledge*, The situation in logic, CSLI lecture notes, Center for the Study of Language and Information, 1989, pp. 201–220.

[Bin90] K. Binmore, *Essays on the foundations of game theory*, Basil Blackwell, Oxford, 1990.

[BIP88] M. E. Bratman, D. J. Israel, and M. E. Pollack, *Plans and resource-bounded practical reasoning*, Computational Intelligence **4** (1988), 349–355.

[Bon96] G. Bonanno, *On the logic of common belief*, Mathematical Logic Quartely **42** (1996), no. 3, 305–311.

[Bra87] M. E. Bratman, *Intentions, plans, and practical reason*, Harvard University Press, Cambridge, MA, 1987.

[Bra90] M. E. Bratman, *What is intention?*, Intentions in Communication (P. R. Cohen, J. L. Morgan, and M. E. Pollack, eds.), MIT Press, 1990, pp. 15–32.

[BS92] K. Binmore and H. S. Shin, *Algorithmic knowledge and game theory*, Knowledge, belief, and strategic interaction (C. Bicchieri and M.-L. Dalla-Chiara, eds.), Cambridge University Press, 1992, pp. 141–154.

[Bur84] J. P. Burgess, *Basic tense logic*, Handbook of Philosophical Logic Volume II — Extensions of Classical Logic (D. Gabbay and F. Guenther, eds.), Reidel, 1984, (Synthese library Volume 164), pp. 89–134.

[Car47] R. Carnap, *Meaning and necessity*, University of Chicago Press, Chicago, 1947.

[Che80] B. Chellas, *Modal logic. An introduction*, Cambridge UP, Cambridge, 1980.

[CL90] P. R. Cohen and H. J. Levesque, *Intention is choice with commitment*, Artificial Intelligence **42** (1990), 213–261.

[CM81] H. H. Clark and C. R. Marshall, *Definite reference and mutual knowledge*, Elements of discourse understanding (A. K. Joshi, B. L. Webber, and I. A. Sag, eds.), Cambridge University Press, 1981.

[Cre70] M. J. Cresswell, *Classical intensional logic*, Theoria **36** (1970), 347–372.

[Cre73] M. J. Cresswell, *Logics and languages*, Methuen, London, 1973.

[Cre80] M. J. Cresswell, *Quotational theories of propositional attitudes*, Journal of Philosophical Logic **9** (1980), 17–40.

[Den87] D. C. Dennett, *The intentional stance*, MIT Press, 1987.

[Ebe74] R. Eberle, *A logic of believing, knowing and inferring*, Synthese **26** (1974), 356–382.

[EDMP91] J. Elgot-Drapkin, M. Miller, and D. Perlis, *Memory, reason, and time: the step-logic approach*, In Philosophy and AI: Essays at the Interface (R. Cummins and J. Pollock, eds.), MIT Press, 1991, pp. 79–104.

[Eme90] E. A. Emerson, *Temporal and modal logic*, Handbook of Theoretical Computer Science (J. van Leeuwen, ed.), vol. B, North-Holland, 1990, pp. 995–1072.

[FG97] S. Franklin and A. Graesser, *Is it an agent, or just a program*, Intelligent Agents III. Proceedings of ATAL-96 (J.P. Müller, M. Wooldridge, and N. Jennings, eds.), LNAI, no. 1193, Springer-Verlag, 1997, pp. 21–36.

[FH88] R. Fagin and J. Y. Halpern, *Belief, awareness and limited reasoning*, Artificial Intelligence **34** (1988), 39–76.

[FHMV95] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi, *Reasoning about knowledge*, MIT Press, Cambridge, Mass., 1995.

[FHV95] R. Fagin, J. Y. Halpern, and M. Y. Vardi, *A nonstandard approach to the logical omniscience problem*, Artificial Intelligence **79** (1995), 203–240.

[Gea92] J. D. Geanakoplos, *Common knowledge*, Journal of Economic Perspectives **6** (1992), 53–82.

[GG93] E. Gillet and P. Gochet, *La logique de la connaissance: Le probleme de l'omniscience logique*, Dialectica **47** (1993), no. 2–3, 143–171.

[Gol87] R. Goldblatt, *Logics of time and computation*, CSLI, Stanford, 1987.

[GR95] M. P. Georgeff and A. S. Rao, *The semantics of intention maintenance for rational agents*, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (Montréal, Québec, Canada), 1995, pp. 704–710.

[Hal87] J. Y. Halpern, *Using reasoning about knowledge to analyze distributed systems*, Annual Review of Computer Science **2** (1987), 37–68.

[Hal95] Steven D. Hales, *Epistemic closure principles*, Southwestern Journal of Philosophy **33** (1995), no. 2, 185–201.

[Har84] D. Harel, *Dynamic logic*, Handbook of Philosophical Logic (D. Gabbay and F. Guenthner, eds.), vol. II, Reidel, 1984, pp. 497–604.

[HC96] G. E. Hughes and M. J. Cresswell, *A new introduction to modal logic*, Routledge, London, 1996.

[Hei95] A. Heifetz, *Common belief in monotonic epistemic logic*, CORE Discussion Paper 9511, Universite Catholique de Louvain, 1995.

[Hin62] J. Hintikka, *Knowledge and belief*, Cornell UP, Ithaca, N.Y., 1962.

[Hin70] J. Hintikka, *Knowledge, belief, and logical consequence*, Ajatus **32** (1970), 32–47.

[Hin75] J. Hintikka, *Impossible possible worlds vindicated*, Journal of Philosophical Logic **4** (1975), 475–484.

[HK91] Z. Huang and K. Kwast, *Awareness, negation and logical omniscience*, Logics in AI, Proceedings JELIA'90 (J. van Eijck, ed.), LNAI, vol. 478, Springer-Verlag, 1991, pp. 282–300.

[HM92] J. Y. Halpern and Y. Moses, *A guide to completeness and complexity for modal logics of knowledge and belief*, Artificial Intelligence **54** (1992), 319–379.

[HMV94] J. Y. Halpern, Y. Moses, and M. Y. Vardi, *Algorithmic knowledge*, Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference (R. Fagin, ed.), Morgan Kaufman, 1994, pp. 255–266.

[Ho93] D. N. Ho, *Ein System der epistemischen Logik*, Philosophie und Logik. Frege-Kolloquien Jena 1989/1991 (W. Stelzner, ed.), Walter de Gruyter, Berlin and New York, 1993, pp. 205–214.

[Ho95] D. N. Ho, *Logical omniscience vs. logical ignorance. On a dilemma of epistemic logic*, Progress in Artificial Intelligence. Proceedings of EPIA'95 (C. P. Pereira and N. Mamede, eds.), LNAI, vol. 990, Springer Verlag, 1995, pp. 237–248.

[Ho97] D. N. Ho, *Reasoning about rational, but not logically omniscient agents*, Journal of Logic and Computation **7** (1997), no. 5, 633–648.

[Ho98] D. N. Ho, *On the epistemic foundations of agent theories*, Intelligent Agents IV. Proceedings of ATAL-97 (M. P. Singh, A. S. Rao, and M. J. Wooldridge, eds.), LNAI, vol. 1365, Springer Verlag, 1998, pp. 275–279.

[Hoc72] M. Hocutt, *Is epistemic logic possible?*, Notre Dame Journal of Formal Logic **13** (1972), 433–453.

[KL88] S. Kraus and D. Lehmann, *Knowledge, belief and time*, Theoretical Computer Science **58** (1988), 155–174.

[Kon86] K. Konolige, *A deduction model of belief*, Pitman Publishing, London, 1986.

[Kor98] F. Koriche, *Approximate reasoning about combined knowledge*, Intelligent Agents IV. Proceedings of ATAL-97 (M. P. Singh, A. S. Rao, and M. J. Wooldridge, eds.), LNAI, vol. 1365, Springer Verlag, 1998, pp. 259–274.

[KT90] D. Kozen and J. Tiurzyn, *Logics of program*, Handbook of Theoretical Computer Science (J. van Leeuwen, ed.), vol. B, North-Holland, 1990, pp. 789–840.

[Len78] W. Lenzen, *Recent work in epistemic logic*, Acta Philosophica Fennica **30** (1978), 1–219.

[Lev84] H. Levesque, *A logic of implicit and explicit belief*, Proceedings AAAI-84 (Austin, TX), 1984, pp. 198–202.

[Lew69] D. Lewis, *Convention*, Harvard Univ. Press, Cambridge, MA, 1969.

[Lew88] D. Lewis, *Desire as belief*, Mind **97** (1988), 323–332.

[Lew96] D. Lewis, *Desire as belief II*, Mind **105** (1996), no. 418, 303–313.

[Lis93] L. Lismont, *La connaissance commune en logique modale*, Mathematical Logic Quarterly **39** (1993), no. 1, 115–130.

[LM94] L. Lismont and P. Mongin, *On the logic of common belief and common knowledge*, Theory and Decision **37** (1994), no. 1, 75–106.

[McC79] J. McCarthy, *Ascribing mental qualities to machines*, Philosophical Perspectives in Artificial Intelligence (Martin Ringle, ed.), Harvester Press, 1979.

[McC90] J. McCarthy, *Formalization of common sense, papers by John McCarthy edited by V. Lifschitz*, Ablex, 1990.

[MH69] J. McCarthy and P. J. Hayes, *Some philosophical problems from the standpoint of artificial intelligence*, Machine Intelligence 4 (B. Meltzer and D. Michie, eds.), Edinburgh University Press, 1969.

[Moo90] R. C. Moore, *A formal theory of knowledge and action*, Readings in Planning (San Mateo, CA) (J. Allen, J. Hendler, and A. Tate, eds.), Morgan Kaufmann Publishers, San Mateo, CA, 1990, pp. 480–519.

[NKP94] M. Nirkhe, S. Kraus, and D. Perlis, *Thinking takes time: a modal active-logic for reasoning in time*, Tech. Report CS-TR-3249, Dept. of Computer Science, Univ. of Maryland, 1994.

[Ran82] V. Rantala, *Impossible worlds semantics and logical omniscience*, Acta Philosophica Fennica **35** (1982), 18–24.

[RG91a] A. S. Rao and M. P. Georgeff, *Asymmetry thesis and side-effect problems in linear time and branching time intention logics*, Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91) (Sydney, Australia), 1991, pp. 498–504.

[RG91b] A. S. Rao and M. P. Georgeff, *Modeling rational agents within a BDI-architecture*, Proceedings of Knowledge Representation and Reasoning (KR&R-91) (R. Fikes and E. Sandewall, eds.), Morgan Kaufmann Publishers, 1991, pp. 473–484.

[Sch72] S. R. Schiffer, *Meaning*, Clarendon, Oxford, 1972.

[Sho93] Y. Shoham, *Agent-oriented programming*, Artificial Intelligence **60** (1993), no. 1, 51–92.

[Sim57] H. Simon, *Models of man*, Macmillan Press, New York, 1957.

[Sin94] M. P. Singh, *Multiagent systems: A theoretical framework for intentions, know-how, and communications*, LNAI, vol. 799, Springer-Verlag, Heidelberg, 1994.

[Sin95] M. P. Singh, *Semantical considerations on some primitives for agent specification*, Intelligent Agents II. Proceedings of ATAL-95 (M. Wooldridge et. al., ed.), LNAI, vol. 1037, Springer-Verlag, 1995, pp. 49–64.

[Ste79] P. Steinacker, *Superschwache Modalkalküle und einige epistemische Anwendungen*, Ph.D. thesis, Universität Leipzig, 1979.

[Ste84] W. Stelzner, *Epistemische Logik*, Akademie-Verlag, Berlin, 1984.

[SW94] H.-S. Shin and Timothy Williamson, *Representing the knowledge of turing machines*, Theory and Decision **37** (1994), no. 1, 125–146.

[Thi93] E. Thijsse, *On total awareness logics*, Diamonds and defaults (M. de Rijke, ed.), Kluwer Academic Publishers Group, 1993, pp. 309–347.

[vB84] J. van Benthem, *Correspondence theory*, Handbook of Philosophical Logic (D. Gabbay and F. Guenthner, eds.), vol. II, Reidel, 1984, pp. 167–248.

[vdHvLM94] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer, *A logic of capabilities (extended abstract)*, Proceedings of the Third International Symposium on the Logical Foundations of Computer Science (LFCS'94) (A. Nerode and Yu.V. Matiyasevich, eds.), LNAI, vol. 813, Springer Verlag, 1994, pp. 366–378.

[vLvdHM94] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer, *Communicating rational agents*, KI-94: Advances in Artificial Intelligence (B. Nebel and L. Dreschler-Fische, eds.), LNAI, vol. 861, Springer Verlag, 1994.

[vLvdHM95a] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer, *Actions that make you change your mind (extended abstract)*, KI-95: Advances in Artificial Intelligence (I. Wachsmuth, C.-R. Rollinger, and W. Brauer, eds.), LNAI, vol. 981, Springer-Verlag, 1995, pp. 185–196.

[vLvdHM95b] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer, *The dynamics of default reasoning (extended abstract)*, Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Proceedings of ECSQARU'95) (C. Froidevaux and J. Kohlas, eds.), LNAI, vol. 946, Springer-Verlag, 1995, pp. 277–284.

[vW63] G. H. von Wright, *Practical inference*, The Philosophical Review **72** (1963), 159–179.

[vW72] G. H. von Wright, *On so-called practical inference*, The Philosophical Review **15** (1972), 39–53.

[Wal92] B. Walliser, *Epistemic logic and game theory*, Knowledge, belief, and strategic interaction (C. Bicchieri and M.-L. Dalla-Chiara, eds.), Cambridge University Press, 1992, pp. 197–225.

[Wan90] H. Wansing, *A general possible worlds framework for reasoning about knowledge and belief*, Studia Logica **49** (1990), no. 4, 523–539.

[WJ95] M. Wooldridge and N. R. Jennings, *Intelligent agents: Theory and practice*, Knowledge Engineering Review **10** (1995), no. 2, 115–152.

[Woo96] M. Wooldridge, *Practical reasoning with procedural knowledge: A logic of BDI agents with know-how*, Practical Reasoning. Proceedings of FAPR'96 (D. Gabbay and H-J. Ohlbach, eds.), LNAI, vol. 1085, Springer Verlag, 1996, pp. 202–213.

[Wut91] K. Wuttich, *Glauben, Zweifel, Wissen. Eine logisch-philosophische studie*, Akademie-Verlag, Berlin, 1991.