

GENERALIZED KRIPKE MODELS FOR EPISTEMIC LOGIC

Frans Voorbraak
 Department of Philosophy, Utrecht University
 P.O. Box 80.126, 3508 TC Utrecht, The Netherlands

ABSTRACT

In this paper a generalization of Kripke models is proposed for systemizing the study of the many different epistemic notions that appear in the literature. The generalized Kripke models explicitly represent an agent's epistemic states to which the epistemic notions refer. Two central epistemic notions are identified: objective (**S5**) knowledge and rational introspective (**KD45**) belief. Their interaction is determined and a notion of justified true belief is explained in terms of them. The logic of this notion of justified true belief is shown to be **S4.2**, which is in accordance with a conjecture by Wolfgang Lenzen. The logic of justified belief is also determined.

1 INTRODUCTION

Although the study of epistemic logic is still rapidly developing, we believe so much is clear: there exist many different notions of knowledge and belief exist which are of interest for research in fields like computer science and AI. For example, when dealing with resource-bounded agents, it is relevant to distinguish between (logically closed) *implicit* and (not necessarily logically closed) *explicit* knowledge/belief. And with respect to the relation between knowledge and belief, one can understand knowledge as true belief, or as justified true belief, belief can be defined as defeasible knowledge or as the conviction of an ideally rational agent, etcetera. The goal of this paper is to obtain a more systematic approach to the formalization of different epistemic notions.

We concentrate on rather idealized notions of knowledge and belief which are closed under logical consequence. We identify two central epistemic notions, which we call *objective knowledge* and *rational belief*, respectively. In Voorbraak (1991) it is argued that if one assumes knowledge to imply rational belief, then one should not use the popular **S5** system of (objective) knowledge. One probably has a notion of knowledge in mind which is closely related to the one most famous among philosophers: justified true belief. The main result of this paper is an explanation of this notion of knowledge in terms of objective knowledge and rational belief, leading to the identification of its logic as **S4.2**, which is the system obtained from **S5** by replacing $\diamond\phi \supset \square\diamond\phi$ (negative introspection) with $\diamond\square\phi \supset \square\diamond\phi$. Although this result may be of some interest in itself, we see it foremost as a demonstration of the utility of our epistemic models.

We essentially use a possible worlds semantics, where the possible worlds are assumed to be complete and consistent. (In a future paper we will consider the issue of allowing partial and

inconsistent worlds in order to capture beliefs which are not necessarily consistent or logically closed.) Similar approaches have already been discussed by several authors. However, our approach not only differs from these earlier ones in the details of the analysis, but also in the way the semantics is given an intuitive justification. We use generalized Kripke models in which the agent's epistemic states are explicitly represented. Different epistemic notions refer to different (aspects of the) epistemic states.

The epistemic states referred to by objective knowledge are called information states and are different from rational belief states. Information states can be ascribed to all information-processing systems, including thermostats, television-receivers, etc. Rational belief states are reserved for rational agents, such as idealized representations of humans. Justified true belief refers to aspects of both information states and rational belief states.

Generalized Kripke models are defined in section 2 below. Sections 3 and 4 identify the notions of objective knowledge and rational belief, respectively, whereas their combination is treated in section 5. (Sections 2-5 consist of revised material of Voorbraak (1991). The proofs of the results mentioned in these sections are omitted in this paper.) Sections 6, 7, and 8 deal with (the logic of) the notions of justified belief and knowledge as justified true belief, and section 9 contains some concluding remarks.

2 GENERALIZED KRIPKE MODELS

Our basic language \mathcal{L} is built up from a set PL of proposition letters (p, q, r,...) and the logical constants \neg and \wedge . We will further use the constants \vee , \supset and \equiv , which are defined as usual. If for all $i \in \{1,2,\dots,n\}$ $[i]$ is a modal operator, then $\mathcal{L}_{[1][2]..[n]}$ is the smallest extension of \mathcal{L} such that if $1 \leq i \leq n$ and $\phi, \psi \in \mathcal{L}_{[1][2]..[n]}$, then $\phi \wedge \psi$, $\neg\phi$, and $[i]\phi \in \mathcal{L}_{[1][2]..[n]}$.

DEFINITION 2.1 A Kripke model for $\mathcal{L}_{[1][2]..[n]}$ is a tuple $M = \langle w_0, W, R_1, R_2, \dots, R_n, \models \rangle$, where $w_0 \in W$, $R_i \subseteq W \times W$, and $\models : W \times \mathcal{L}_{[1][2]..[n]} \rightarrow \{T, F\}$ such that

- $\models(w, \phi \wedge \psi) = T$ iff $\models(w, \phi) = \models(w, \psi) = T$
- $\models(w, \neg\phi) = T$ iff $\models(w, \phi) = F$
- $\models(w, [i]\phi) = T$ iff $\forall w' \in W (wR_i w' \Rightarrow \models(w', \phi) = T)$.

We usually write $w \models \phi$ instead of $\models(w, \phi) = T$ and we define $M \models \phi$ iff $w_0 \models \phi$. w_0 is called the *designated world* of M .*

* The use of Kripke models with designated worlds has several advantages. For example, the definition of $\Gamma \models_S \phi$ simply becomes: every S-model of Γ is also a model of ϕ . It is not necessary to quantify over the worlds in the models and if we were to allow inconsistent or incomplete worlds to model less idealized notions of belief, then we would not have to restrict the quantification to complete and consistent worlds, as in Levesque (1984): we simply assume the designated worlds to be complete and consistent (being possible *real* worlds rather than being imagined). The (more usual) Kripke models without designated worlds can be considered to be equivalence classes of the models with designated world and have of course also some advantages, since it is sometimes convenient to abstract from the particular choice of a designated world.

DEFINITION 2.2 A *generalized Kripke model* for the language $\mathcal{L}_{[1][2]..[n]}$ is a tuple $M = \langle w_0, W, \langle \Sigma_1, \Sigma_2, \dots, \Sigma_m \rangle, \langle \sigma_1, \sigma_2, \dots, \sigma_m \rangle, \langle F_1, F_2, \dots, F_m \rangle, \models \rangle$, where $w_0 \in W$, Σ_i is a non-empty set, $\sigma_i : W \rightarrow \Sigma_i$, F_i is a set of functions with domain Σ_i , and $\models : W \times \mathcal{L}_{[1][2]..[n]} \rightarrow \{T, F\}$ s.t.

- $\models(w, \phi \wedge \psi) = T$ iff $\models(w, \phi) = \models(w, \psi) = T$
- $\models(w, \neg \phi) = T$ iff $\models(w, \phi) = F$
- $\models(w, [i]\phi)$ functionally depends on w , $\models_{\uparrow W \times \{\phi\}}$, $\langle \sigma_1, \dots, \sigma_m \rangle$, and $\langle F_1, \dots, F_m \rangle$.

Σ_i denotes a set of epistemic states, $\sigma_{[i]}$ associates with every world an element of Σ_i , and for every epistemic state $S \in \Sigma_i$ and function $f \in F_i$ $f(S)$ denotes an aspect of the epistemic state S . Elements of F_i are called *projection functions*, since they extract information out of an epistemic state by "projecting" the epistemic state onto some aspect of the state. An example of a projection function is the function $\|\cdot\| : \Sigma_i \rightarrow \wp W$ given by $\|S\| = \{w \in W \mid w \text{ is possible according to } S\}$. Another example is the function Π such that $\Pi(S)$ yields a probability measure on $\wp \|S\|$ giving the probabilities which according to S have to be assigned to the (sets of) possible worlds.

In general, we will only exhibit those projection functions which are mentioned in the valuation clauses of the modalities or which are used to single out a particular class of generalized Kripke models. In the above definition we do not require that $m = n$, since a modal operator may refer to more than one epistemic state and there may be different modal operators referring to the same epistemic states.

Generalized Kripke models generalize Kripke models in two respects: (1) epistemic states occur as atomic entities in the generalized models; they are no longer (indirectly) represented by means of sets of possible worlds and (2) in the generalized models modal operators may have nonstandard valuation clauses. Nevertheless, a class of generalized Kripke models will often turn out to be equivalent to a class of Kripke models, so one can often invoke well-known results and methods to obtain soundness and completeness results.

In the following section, an example of the use of non-standard valuation clauses will be given. Our justification for the explicit representation of epistemic states in the generalized Kripke models is that it helps to systemize the study of the relations between various epistemic modalities. For example, the two central notions in this paper - objective knowledge and rational belief - refer to different epistemic states, and are therefore not as strongly related as some authors seem to think, whereas some (strongly related) notions of knowledge and belief might be considered to refer to different aspects of the same epistemic state. In general, we believe the generalized Kripke models to be useful in bringing the study of epistemic logic to a higher level of 'informal rigour' (in the sense of Kreisel (1967)).

3 OBJECTIVE KNOWLEDGE

Let K_α be the operator with the following intuitive meaning:

$K_\alpha \phi$: agent α objectively knows that ϕ , i.e., ϕ is the case (in every world that is possible) given the information available to α .

Since we will restrict ourselves to the one agent case, we will omit the subscript. The notion of objective knowledge applies to any agent which is capable of processing information, in the sense of Dretske (1981)*; we do not require the agent to consider the (im)possibility of some worlds. The logic of this notion of objective knowledge will be denoted by **OK**. Closely following the above intuitive explanation of the meaning of the operator **K**, we propose the following semantics for **K**, which can easily be seen to lead to the modal logic **S5** as the logic for objective knowledge:

DEFINITION 3.1 An *objective knowledge model*, or **OK model**, is a generalized Kripke model $M = \langle w_0, W, \Sigma_K, \kappa, \models \rangle$, where $w \models K\phi \Leftrightarrow \forall w' \in W (\kappa(w') = \kappa(w) \Rightarrow w' \models \phi)$. Alternatively one could add the projection function $\|\cdot\|_K$, given by $\|\kappa(w)\|_K = \{w' \mid \kappa(w) = \kappa(w')\}$ and define $w \models K\phi \Leftrightarrow \forall w' \in \|\kappa(w)\|_K w' \models \phi$. Elements of Σ_K are called *information states*.

PROPOSITION 3.2 $\forall \phi \in \mathcal{L}_K (\models_{S5} \phi \Leftrightarrow \models_{OK} \phi)$.

Albeit the introspection axioms ($K\phi \supset KK\phi$ and $\neg K\phi \supset K\neg K\phi$) are valid in **OK**, we did not assume our agents to have some kind of privileged access to or a perfect awareness of their epistemic attitudes. In fact, the logic of **K** is determined even though the notion of information state is almost completely unspecified. The equivalence of **S5** models and **OK** models does of course not show that information states can be equated with sets of possible worlds, since the equivalence depends for example on the fact that that the language \mathcal{L}_K does not admit one to exploit probabilistic information which may be contained in the information states.

The above method of identifying the logic of objective knowledge as **S5** is a generalization of the method by which the logic of the knowledge of processors in a distributed system is identified to be **S5**. In the latter case, one takes the information state of a processor to be identical to the internal state of the processor. (See for example Halpern and Moses (1984).)

4 RATIONAL (INTROSPECTIVE) BELIEF

In this section we consider the operator B_α with the following intuitive meaning:

$B_\alpha\phi$: agent α rationally believes (is rationally convinced) that ϕ , i.e., ϕ is valid in every world that is considered possible by α .

Again we usually omit the subscript. The above notion of belief is a very strong one. In fact, if $B_\alpha\phi$, then α will be inclined to express his belief by saying that he *knows* that ϕ . If an agent says that he (only) *believes* that ϕ , then most likely a weaker notion of belief is used. Some possible weaker notions will be treated in a future paper.

* Dretske's notion of knowledge additionally requires the presence of a higher-order intentional belief state and consequently differs considerably from the notion of objective knowledge.

Let $\beta(w)$ be the rational belief state (of an agent α) in w and let $\|\beta(w)\|_{\beta}$ denote the set of worlds which are considered possible (by α) in w . (The subscripts of $\|\cdot\|$ will usually be omitted.) Rational belief states are assumed to satisfy the following two conditions:

CONSISTENCY: For every rational belief state $\beta(w)$ $\|\beta(w)\| \neq \emptyset$ (I)

INTROSPECTION: For every rational belief state $\beta(w)$ ($w' \in \|\beta(w)\| \Rightarrow \beta(w') = \beta(w)$) (II)

The consistency-condition derives from the assumed rationality of the agent: a rational agent will not adhere to an inconsistent set of beliefs, but will revise this set to a consistent one. The introspection-condition derives from the assumption that a rational agent is fully aware of his own belief state. Hence in every world he considers to be a possible candidate of the actual world he must have the same belief state. A rational agent is not infallible, hence we do *not* require $w \in \|\beta(w)\|$. Some other conditions on rational belief states (concerning the relation between rational belief states and information states) are treated in the following section.

We do not equate a rational agent's belief state S with $\|S\|$, since we do not want to exclude the possibility that the agent has e.g. beliefs concerning the relative likelihood of subclasses of $\|S\|$. **RIB**, the logic of rational introspective belief has the following semantics:

DEFINITION 4.1 A *rational belief model*, or **RIB model**, is a generalized Kripke model $M = \langle w_0, W, \Sigma_B, \beta, \|\cdot\|, \models \rangle$, which satisfies conditions I and II, and $w \models B\phi \Leftrightarrow \forall w' \in \|\beta(w)\| w' \models \phi$. Elements of Σ_B are called (*rational*) *belief states*.

Just as in the case of the **OK** models, the class of **RIB** models is equivalent to a class of very familiar Kripke models, namely the Kripke models for **KD45**.

PROPOSITION 4.2 $\forall \phi \in \mathcal{L}_B (\models_{KD45} \phi \Leftrightarrow \models_{RIB} \phi)$.

5 COMBINING OBJECTIVE KNOWLEDGE AND RATIONAL BELIEF

Determining how **K** and **B** interact can be done in a systematic way: Consider a world w . Describe what is known or can be assumed of (1) the belief state in worlds relevant for the valuation of $K\phi$ and (2) the information state in worlds relevant for the valuation of $B\phi$. In other words, describe what can be assumed of $\beta(w')$ if it is known that $\kappa(w) = \kappa(w')$ and of $\kappa(w')$ if it is known that $w' \in \|\beta(w)\|$.

The information objectively available to a rational introspective agent surely contains all information about his own rational belief state. Hence if $\kappa(w) = \kappa(w')$ (or $w' \in \|\kappa(w)\|$), then $\beta(w')$ is completely determined: $\beta(w') = \beta(w)$.^{*} Especially in this information age, it cannot be assumed that a rational agent represents all information objectively available to him in his belief

* It is tempting to think of belief states as aspects of information states. However, one must remember that rational belief states cannot be ascribed to every information-processing system.

state. Thus we do not have $w' \in \|\beta(w)\| \Rightarrow \kappa(w') = \kappa(w)$. However, if $w' \in \|\beta(w)\|$, then it is reasonable to assume that $\|\kappa(w')\| \subseteq \|\beta(w)\|$, since a world which is considered(!) to be objectively possible from a world believed to be a possible candidate for the actual world should itself be believed to be a possible candidate for the actual world. Hence we arrive at the following assumptions on the relation between information states and rational belief states*:

$$\text{OBJECTIVITY: } \forall w, w' (w' \in \|\kappa(w)\| \Rightarrow \beta(w') = \beta(w)) \tag{III}$$

$$\text{BELIEVED SOUNDNESS: } \forall w, w' (w' \in \|\beta(w)\| \Rightarrow \|\kappa(w')\| \subseteq \|\beta(w)\|) \tag{IV}$$

The semantics for **OK&RIB**, the system in which objective knowledge and rational introspective belief are combined, becomes:

DEFINITION 5.1 An *objective knowledge & rational belief model*, or **OK&RIB model**, is a generalized Kripke model $M = \langle w_0, W, \langle \Sigma_K, \Sigma_B \rangle, \langle \kappa, \beta \rangle, \langle \|\cdot\|_K, \|\cdot\|_B \rangle, \models \rangle$ for \mathcal{L}_{KB} such that $\kappa, \beta, \|\cdot\|_K, \|\cdot\|_B$, and the valuation clauses of **K** and **B** are as before and conditions I - IV are satisfied.

OK&RIB models correspond to Kripke models for \mathcal{L}_{KB} such that R_K is an equivalence relation, R_B is serial, transitive, and euclidean, and $R_B R_K \subseteq R_B \supseteq R_K R_B$. (Since R_K is reflexive, the inclusions may be replaced by equations.) Hence **OK&RIB** is the system which combines **S5 K** and **KD45 B** by means of $B\phi \supset BK\phi$, $B\phi \supset KB\phi$. In this system, $B\phi$, $KB\phi$, and $BK\phi$ are equivalent and we have: $K\neg B\phi \equiv \neg B\phi$, $B\neg\phi \supset B\neg K\phi$, and $B\neg K\phi \supset \neg B\phi$.

6 KNOWLEDGE WHICH IMPLIES RATIONAL BELIEF

Objective knowledge and rational belief seem to correspond with the nowadays predominant notions of knowledge and belief. Nevertheless, **OK&RIB** differs in a number of respects from the system - let us call it **KL** - which was proposed in Kraus and Lehmann (1986) and thoroughly studied from a technical point of view in van der Hoek (1991). The main difference is that **KL** has $K\phi \supset B\phi$ among its axioms, whereas this formula, which corresponds to $w' \in \|\beta(w)\| \Rightarrow \kappa(w') = \kappa(w)$, is not derivable in **OK&RIB**. In Voorbraak (1991) we extensively argue that the popular **S5** notion of knowledge should not imply rational (**KD45**) belief. However, one frequently seems to use a notion of knowledge which *does* imply belief. For example, in the philosophical literature knowledge is often equated with justified true belief. Such a notion of knowledge K_j is obviously different from our notion of objective knowledge and the justification given in section 3 for taking **S5** as the logic of **K** does not apply to K_j .

In Lenzen (1979) it is argued that the logic of K_j is at least as strong as **S4.2** ($= \mathbf{S4} + \diamond\Box\phi \supset \Box\diamond\phi$) and at most as strong as **S4.4** ($= \mathbf{S4} + \phi \supset (\diamond\Box\phi \supset \Box\phi)$). The latter system is

* In Voorbraak (1991) we also considered (and rejected) believed completeness: $w' \in \|\beta(w)\| \Rightarrow \|\kappa(w')\| \supseteq \|\beta(w)\|$, which might be appropriate for an (overconfident) agent who believes to have fully exploited all the information objectively available to him. The principle corresponding to this assumption is $\neg B\phi \supset B\neg K\phi$.

shown to be the logic of true rational (KD45) belief. Lenzen conjectures that the logic of K^j is **S4.2**, which is supported by the results of the following sections. An obvious definition of K^j is $K^j\phi = B\phi \wedge J\phi \wedge \phi$, where $J\phi$ denotes something like 'the agent is justified in believing that ϕ '. However, the following example shows that this definition leads to problems:

EXAMPLE 6.1 ($K^j\phi \neq B\phi \wedge J\phi \wedge \phi$) Assume that an agent believes that p , but he is not justified in believing that p , and that he is justified in believing that q , but he does not believe that q . Finally, assume that q is true. Then he believes $p \vee q$ and, under some reasonable assumptions, he is justified in believing $p \vee q$, and of course $p \vee q$ is true. Yet, intuitively, his belief in $p \vee q$ is not justified, since he believes $p \vee q$ "for the wrong reason".

A better reading of $J\phi$ is 'the agent's belief that ϕ is justified' or 'the agent justifiably believes that ϕ '. Since under this interpretation $J\phi$ implies $B\phi$, it seems appropriate to use the more suggestive notation $B^j\phi$ instead of $J\phi$ and define $K^j\phi = B^j\phi \wedge \phi$. Since it is reasonable to assume that true beliefs are not necessarily justified, B^j cannot be equated with B . We will also assume that justified beliefs are not necessarily true. This assumption makes a further adjustment of the definition of K^j necessary:

EXAMPLE 6.2 ($K^j\phi \neq B^j\phi \wedge \phi$) Assume that an agent justifiably believes that p , and that he does not believe that q . Further, assume that $\neg p \wedge q$ is true. Then he believes $p \vee q$ and, under some reasonable assumptions, he justifiably believes that $p \vee q$, and of course $p \vee q$ is true. Yet, intuitively, he does not know that $p \vee q$, since he believes $p \vee q$ "for the wrong reason". (This is essentially 'Case II' of Gettier (1963). See Lenzen (1978) for an overview of some of the many reactions to Gettier's paper.)

Hence we will interpret $K^j\phi$ as $B^j\phi \wedge$ ' ϕ is true for the same reasons why ϕ is justifiably believed'. Since the second conjunct implies the first, we now have completely left the idea of defining knowledge as a conjunction of more simple expressions. This is in line with the analysis of knowledge given in Lehrer (1990), where four conditions for knowledge are given such that one of them (undefeated justified acceptance) implies the other three (truth, acceptance and complete justification). In the following section we give a semantics for B^j and K^j , building on the previously defined generalized Kripke models for objective knowledge K and rational belief B .

7 A SEMANTICS FOR JUSTIFIED (TRUE) BELIEF

Suppose a rational agent believes that ϕ . It is reasonable to say that this belief is justified iff the agent has some good reasons for it, where a reason is called good iff it is in some sense supported by the information which is objectively available to the agent. In this paper, a reason is interpreted by a set S of possible worlds and is called good whenever S contains (a counterpart to) every unexceptional world possible given the information objectively available to the agent. Hence a good reason for ϕ does *not* necessarily imply the truth of ϕ , but a reason is called good if it only

discards *exceptional* objective possibilities. Let us give an example before making these ideas more precise.

EXAMPLE 7.1 When Bill approaches Jack's house he has two reasons for believing that Jack is at home: Jack's car is parked in front of the house and the tones of John Coltrane's "A Love Supreme" come from the house. Both are good reasons, since Jack almost never leaves the house without his car and his wife and children do not share Jack's love for jazz. Each of these reasons may on its own already be sufficient to say that Bill knows that Jack is at home, provided that this is actually true and for that reason. Bill's belief does not count as knowledge if Jack is not at home, or if Jack is at home, but he has been using his wife's car all day, since his own car is defect, and Coltrane is played by his children (by mistake).

Hence a belief supported by good reasons is not necessarily true, and even if it is true it does not necessarily constitute knowledge. Further it may be noted that the notion of a good reason (and therefore that of knowledge) is not absolute. The mentioned reasons may normally be good enough, but will probably not suffice in case Jack's presence is a matter of life and death (for example, if only Jack is capable of restraining his dog Rambo when a visitor arrives). In that case, not only should Bill be more sceptical in deciding whether the reasons sufficiently support his beliefs, but it will also be more difficult for reasons to be good enough to *justify* Bill's belief.

Let υ denote a function which assigns to each information state $\kappa(w)$ the set of unexceptional worlds possible given the information objectively available to the agent in w . We only require $\upsilon(\kappa(w))$ to be a subset of the set $\|\kappa(w)\|$ of objectively possible worlds. In particular, we do not exclude that all objectively possible worlds are unexceptional nor do we exclude that they are all exceptional. Depending on the application some further requirements on υ can be formulated, resulting in a stricter interpretation of the concept of a good reason. However, the logic of KJ can be determined without further specification of υ . In other words, KJ will represent several notions of justified true belief which all have the same logic.

The reasons underlying a rational belief state $\beta(w)$ can be divided into specific, a posteriori reasons and general, a priori ones. The specific reasons are assumed to be represented by a collection $\rho(\beta(w))$ of sets of worlds. An element S of $\rho(\beta(w))$ corresponds with a reason for believing that the (real) world must be a member of S . Typically, an element S of $\rho(\beta(w))$ will be the set of worlds satisfying some proposition ϕ , which represents or is inferred from some specific evidence. A rational (introspective) agent can also put *a priori* constraints on worlds: for example, in every possible world the belief state of the agent has to be equal to his belief state in the present world. One can think of $\|\beta(w)\|$ as the set of worlds obtained from a set of a priori possible worlds by deleting those worlds which do not satisfy some specific constraints. In other words, if we write A_w for the set of worlds which are a priori believed to be possible in w , then we can require that $\|\beta(w)\| = A_w \cap \bigcap \rho(\beta(w))$.

We now proceed to give a definition of A_w . As stated before, we assume that $w' \in A_w \Rightarrow \beta(w') = \beta(w)$. Further, the worlds which are believed to be objectively possible in $w' \in A_w$ should be a priori possible in w and should not be excluded by the (specific) reasons according to which w' is believed to be possible. Hence if we define $CR_w(w') = \bigcap \{S \in \rho(\beta(w)) \mid w' \in S\}$ (=

the conjoined specific reasons in w for believing that w' might be the actual world), then $w' \in A_w \Rightarrow \|\kappa(w')\| \subseteq CR_w(w')$. More general, we require for every $\Sigma \subseteq \rho(\beta(w))$ that $A_w \cap \cap \Sigma$ satisfies the conditions we have previously formulated for $\|\beta(w)\|$:

GENERALIZED CONSISTENCY: $\forall w \forall \Sigma \subseteq \rho(\beta(w)) \ A_w \cap \cap \Sigma \neq \emptyset$

GENERALIZED INTROSPECTION: $\forall w, w' \forall \Sigma \subseteq \rho(\beta(w)) \ (w' \in A_w \cap \cap \Sigma \Rightarrow \beta(w') = \beta(w))$

GEN. BELIEVED SOUNDNESS: $\forall w, w' \forall \Sigma \subseteq \rho(\beta(w)) \ (w' \in A_w \cap \cap \Sigma \Rightarrow \|\kappa(w')\| \subseteq A_w \cap \cap \Sigma)$

GENERALIZED CONSISTENCY follows immediately from CONSISTENCY and $\|\beta(w)\| = A_w \cap \cap \rho(\beta(w))$. In the generalizations of INTROSPECTION and BELIEVED SOUNDNESS one can drop the quantification over subsets of $\rho(\beta(w))$ and replace $\cap \Sigma$ with $CR_w(w')$ or even with $CR(w') = CR_w(w')$. A_w is defined as the set of worlds satisfying the above conditions. More precisely, A_w is the largest subset of W such that $w' \in A_w \Rightarrow \beta(w') = \beta(w) \ \& \ \|\kappa(w')\| \subseteq A_w \cap CR(w')$. The self-reference in this definition is harmless, since A_w can be given the following alternative noncircular characterization: $A_w = \{u \in W \mid \beta(u) = \beta(w) \ \& \ \forall v \in \|\kappa(u)\| \ \|\kappa(u)\| \subseteq CR(v)\}$. Since A_w is determined by κ , β , and ρ , the requirement $\|\beta(w)\| = A_w \cap \cap \rho(\beta(w))$ is a constraint on ρ .

It remains to define which reasons underlying a belief state count as *good* reasons. The a priori reasons are quite trustworthy, but among the specific reasons there may be some deriving from unreliable sources of evidence. An obvious first attempt to single out the good specific reasons is to require them to contain all unexceptional worlds. However, we need a somewhat more complex definition: a specific reason $S \in \rho(\beta(w))$ is called a *good* reason iff the set of a priori possible counterparts to each unexceptional world is not empty and is contained in S . An argument for this modification is given below, but we first define the notion of counterpart:

If u is considered a priori possible in w , then u is its own and only counterpart. Otherwise, u' is called a counterpart of u (in w) whenever u' and u satisfy the same nonmodal formulas, $\beta(u) = \beta(u')$, and each unexceptional objectively possible world in u has some counterpart which is objectively possible in u' . Hence if $uR_w u'$ denotes that u' is a counterpart of u (in w), then $\{u' \mid uR_w u'\}$ contains all the worlds which are to some degree equivalent to u (in w). More precisely, R_w is the union of $\{\langle x, x \rangle \mid x \in A_w\}$ and the largest subset of $(W - A_w) \times W$ s.t. $xR_w y \Rightarrow \forall \varphi \in \mathcal{L} \ (x \models \varphi \Leftrightarrow y \models \varphi)$, $\beta(x) = \beta(y) \ \& \ \forall u \in v(\kappa(x)) \ \exists v \in \|\kappa(y)\| \ uR_w v$. Again, the self-reference can be disposed of: $uR_w v$ can be replaced by $\forall \psi \in \mathcal{L} \ (u \models \psi \Leftrightarrow v \models \psi)$.

Counterparts to possible worlds are in general only required to be *partially* equivalent to those worlds. In particular, we do not require that a counterpart u' to some world $u \notin A_w$ has the same information state as u . The rationale behind this (and behind the rejection of the first attempt to define good reasons) is that the truth of $K\varphi$ is determined by an 'all or nothing' clause, which does not correspond well to the interpretation of a good reason as being *usually* truth-implying: as soon as φ is false in one (perhaps very exceptional) objective possibility, then $K\varphi$ is false in all objective possibilities. Since $B\varphi \equiv BK\varphi$, a good reason for believing that φ should also be a good reason for believing that φ is objectively known. And indeed, $B^j\varphi \equiv B^jK\varphi$ turns out to be valid in the semantics proposed below.

The set of good reasons $G_w \subseteq \rho(\beta(w))$ is given by $G_w = \{S \in \rho(\beta(w)) \mid \forall u \in v(\kappa(w)) \ \emptyset \neq A_w \cap \{u' \mid uR_w u'\} \subseteq S\}$. Notice that the notion of a good reason is not absolute: it depends

on how the concept of unexceptional world is filled in. We say that in a world w an agent justifiably believes that φ iff φ is true in every world w' which is considered a priori possible in w and which is an element of every good reason $S \in G_w$. We say that φ is justifiably *known* in w iff φ is true in w and in every world w' which is considered a priori possible in w and which is an element of every good reason $S \in G_w$ such that $w \in S$. The italicized condition is added to guarantee that the reasons for believing that φ correspond to the actual reasons for φ being true. Hence we arrive at the following definition:

DEFINITION 7.2 A *justified true belief model*, or **JTB model**, is a generalized Kripke model $M = \langle w_0, W, \langle \Sigma_K, \Sigma_B \rangle, \langle \kappa, \beta \rangle, \langle \{ \|\cdot\|_K, \nu \}, \{ \|\cdot\|_B, \rho \} \rangle, \models \rangle$ for $\mathcal{L}_{KBK^jB^j}$, where

- $\kappa, \beta, \|\cdot\|_K, \|\cdot\|_B$, and the valuation clauses of K and B are as before
- conditions I - IV and the generalizations of II and IV are satisfied
- $\nu : \Sigma_K \rightarrow \wp W$ such that $\nu(\kappa(w)) \subseteq \|\kappa(w)\|_K$
- $\rho : \Sigma_B \rightarrow \wp \wp(W)$, such that $\|\beta(w)\|_B = A_w \cap \cap \rho(\beta(w))$.
- $w \models B^j\varphi$ iff $\forall w' \in A_w \cap \cap G_w w' \models \varphi$
- $w \models K^j\varphi$ iff $\forall w' \in (A_w \cap \cap T_w) \cup \{w\} w' \models \varphi$, where $T_w = \{S \in G_w \mid w \in S\}$.

The examples below show that in our semantics justified belief (B^j) is not the same as true belief, and justified true belief (K^j) is not the same as justified belief which happens to be true.

EXAMPLE 7.3 ($B^j\varphi \neq B\varphi \wedge \varphi$) Assume that $PL = \{p, q\}$ and consider the **JTB** model $M = \langle w_0, W, \langle \Sigma_K, \Sigma_B \rangle, \langle \kappa, \beta \rangle, \langle \{ \|\cdot\|_K, \nu \}, \{ \|\cdot\|_B, \rho \} \rangle, \models \rangle$, where $W = \{w_0, w_1, w_2\}$, $w_0 \models \neg p \wedge q$, $w_1 \models p \wedge \neg q$, $w_2 \models p \wedge q$, $\Sigma_K = \Sigma_B = \wp W$, $\|\cdot\|_K = \|\cdot\|_B = \text{id}_{\wp W}$, $\kappa(w_0) = \nu(\kappa(w_0)) = \{w_0\}$, $\forall i \in \{1, 2\} \kappa(w_i) = \nu(\kappa(w_i)) = \{w_1, w_2\}$, and $\forall i \in \{0, 1, 2\} \beta(w_i) = \{w_1, w_2\}$, $\rho(\beta(w_i)) = \{\{w_1, w_2\}\}$. Then $M \models B(p \vee q)$, $M \models p \vee q$, but not $M \models B^j(p \vee q)$.

EXAMPLE 7.4 ($K^j\varphi \neq B^j\varphi \wedge \varphi$) Assume that $PL = \{p, q\}$ and consider the **JTB** model $M = \langle w_0, W, \langle \Sigma_K, \Sigma_B \rangle, \langle \kappa, \beta \rangle, \langle \{ \|\cdot\|_K, \nu \}, \{ \|\cdot\|_B, \rho \} \rangle, \models \rangle$, where $W = \{w_0, w_1, w_2\}$, $w_0 \models \neg p \wedge q$, $w_1 \models p \wedge \neg q$, $w_3 \models p \wedge \neg q$, $\Sigma_K = \Sigma_B = \wp W$, $\|\cdot\|_K = \|\cdot\|_B = \text{id}_{\wp W}$, $\forall i \in \{0, 1\} \kappa(w_i) = \{w_0, w_1\}$, $\nu(\kappa(w_i)) = \{w_1\}$, $\kappa(w_2) = \nu(\kappa(w_2)) = \{w_2\}$, and $\forall i \in \{0, 1, 2\} \beta(w_i) = \{w_2\}$, $\rho(\beta(w_i)) = \{\{w_1, w_2\}\}$. Then $M \models B^j(p \vee q)$, $M \models p \vee q$, but not $M \models K^j(p \vee q)$.

8 THE LOGIC OF JUSTIFIED (TRUE) BELIEF

The following lemma lists some useful properties of notions defined in the previous section.

LEMMA 8.1

- (i) $w' \in A_w \Rightarrow (A_w = A_{w'} \ \& \ R_w = R_{w'})$.
- (ii) $w' \in A_w \cap \cap T_w \Rightarrow (\|\kappa(w')\| \subseteq A_w \cap \cap T_w \ \& \ A_{w'} \cap \cap T_{w'} \subseteq A_w \cap \cap T_w)$.

Using the above properties, which have straightforward proofs, it is not difficult to check that in **JTB** models the following formulas are valid: $K^j\phi \supset \phi$, $K^j\phi \supset K^jK^j\phi$, $\neg K^j\neg K^j\phi \supset K^j\neg K^j\neg\phi$. Since also K^j -distribution and K^j -necessitation are valid, the logic of K^j is at least **KT4G** or **S4.2**. In fact, the logic of K^j is exactly **S4.2**, since if one restrict the language to \mathcal{L}_{K^j} , then for every **S4.2** model there exists an equivalent **JTB** model. [Reminder: **S4.2** models are reflexive, transitive and incestual ($\forall x,y,y' (xRy \wedge xRy' \Rightarrow \exists z (yRz \wedge y'Rz))$).]

LEMMA 8.2 The existence of a finite **S4.2** countermodel to $\phi \in \mathcal{L}_{K^j}$ implies the existence of a **JTB** countermodel to ϕ .

Proof. Let $M = \langle w_0, W, R, \models \rangle$ be a finite **S4.2** countermodel to ϕ , which is generated by w_0 . Define $M' = \langle w_0, W, \langle \Sigma_K, \Sigma_B \rangle, \langle \kappa, \beta \rangle, \langle \{ \parallel \cdot \parallel_K, \nu \}, \{ \parallel \cdot \parallel_B, \rho \} \rangle, \models' \rangle$, where $\Sigma_K = \Sigma_B = \wp W$, $\parallel \cdot \parallel_K = \parallel \cdot \parallel_B = \text{id}_{\wp W}$, $\forall w \in W \kappa(w) = \{ w' \in W \mid wRv \Rightarrow w'Rv \}$, $\beta(w) = \{ w' \in W \mid w'Rv \Rightarrow vRw' \}$, $\nu(\kappa(w))$ is any subset of $\kappa(w)$, $\rho(\beta(w)) = \{ \{ v \mid uRv \} \mid u \in W \}$, and $\forall p \in PL (w \models' p \text{ iff } w \models p)$. (See fig. 1 for a picture of κ , β , and ρ against the background of a typical **S4.2** model.) Then it is entirely routine to check that M' is a **JTB** model in which ϕ is not valid.

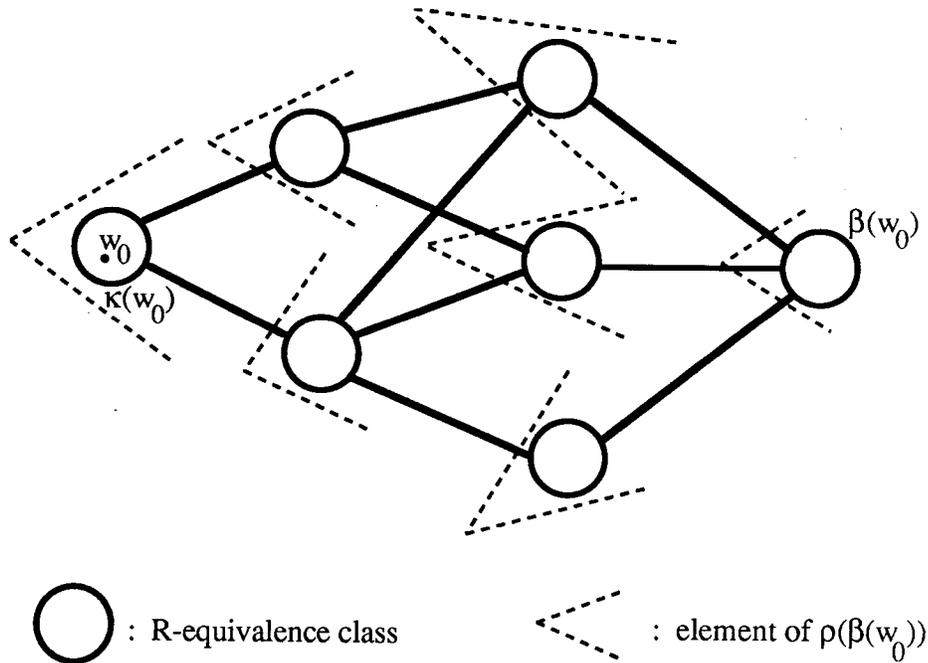


Figure 1: Building a **JTB** model out of a **S4.2** model. Fat lines denote the **S4.2** model.

By the finite model property of **S4.2**, we may conclude:

PROPOSITION 8.3 $\forall \phi \in \mathcal{L}_{K^j} (\models_{S4.2} \phi \Leftrightarrow \models_{JTB} \phi)$.

This proposition supports Wolfgang Lenzen's conjecture that the logic of justified true belief is **S4.2**. (Lenzen (1978,1979).) Notice that the proposition does not depend on the interpretation of 'unexceptional world', since the proof of lemma 8.2 goes through for any $v(\kappa(w)) \subseteq \|\kappa(w)\|$. In fact, the only role v plays is that of forcing the validity of $K^j\phi \supset B^j\phi$.

For the logic of B^j , consider the following schemes

T' $\Box(\Box\phi \supset \phi)$

(T \vee 5)' $\Box((\Box\phi \supset \phi) \vee (\neg\Box\phi \supset \Box\neg\Box\phi))$

N $(\Diamond\phi \wedge \Box(\Diamond\phi \supset \phi_0) \wedge \Diamond\psi \wedge \Box(\Diamond\psi \supset \psi_0)) \supset (\Diamond\chi \supset \Diamond(\phi_0 \wedge \psi_0 \wedge \Diamond\chi))$ [$\phi_0, \psi_0 \in \mathcal{L}$]

We define **JB** as **KD4GN(T \vee 5)'** and **JB'** as **KD4GNT'**. It can be shown that the logic of B^j is **JB**, and if one additionally assumes that unexceptional objective possibilities always exist, i.e., $\forall w \in W v(\kappa(w)) \neq \emptyset$, then the logic is **JB'**. **JB** and **JB'** are both incomplete systems, since any frame of **JB** or **JB'** has to satisfy $\Box(\Box\phi \vee \Box\psi) \supset (\Box\phi \vee \Box\psi)$, which is not valid in **JB** models nor in **JB'** models. Nevertheless, interesting (non-trivial) Kripke semantics for **JB** and **JB'** can be given. (See the Appendix.)

9 CONCLUSION

We have illustrated the use of generalized Kripke models for epistemic logic, by deriving the logic of some important epistemic notions: objective knowledge, rational belief, and justified (true) belief. In Voorbraak (1991) it is argued that the **S5** notion of objective knowledge does not imply rational (**KD45**) belief. Authors who believe that knowledge implies rational belief probably have a notion of knowledge as justified true belief in mind.

In this paper the logic of justified true belief is identified as **S4.2**, in accordance with a conjecture of Wolfgang Lenzen. Somewhat surprisingly, the logic of justified belief turns out to be an incomplete modal logic and thus more complex from a technical point of view. Since the models for justified true belief K^j are also models for objective knowledge K and rational belief B , it is reasonably straightforward to extend the analysis to the interaction between K^j , K , and B . For example, the interaction between K^j and B can be shown to be governed by the schemes $K^j\phi \supset B\phi$, $B\phi \supset BK^j\phi$, and $\neg B\phi \supset B\neg K^j\phi$. Equivalently, one could add the definition $B\phi \equiv \neg K^j\neg K^j\phi$ to **S4.2** for K^j . This definition of belief in terms of a notion of knowledge at least as strong as **S4.2** is already proposed in Lenzen (1979). (Hence Shoham and Moses (1989) were not the first to consider a definition of belief in terms of knowledge.)

As we see it, the important advantages of the use of generalized Kripke models are:

- The informal considerations underlying the formalization of epistemic notions have to be made more precise.
- Extending the analysis to notions referring to new epistemic states leads to extended models (having more Σ_i); extending the analysis to notions referring to new aspects of old epistemic states leads to more refined models (having more projection functions).

In short, generalized Kripke models form promising instruments for a systematic study of the many interesting epistemic notions.

10 APPENDIX

In this appendix we show, assuming a working knowledge of modal logic up to the level of Chellas (1980), that the logic of **BJ** is **JB**, and if one additionally assumes that unexceptional objective possibilities always exist, i.e., $\forall w \in W \ v(\kappa(w)) \neq \emptyset$, then the logic is **JB'**. Using the properties of lemma 10.1 below it is easy to see that **JB** and **JB'** form lower bounds for the respective logics.

LEMMA 10.1

- (i) $w' \in A_w \cap \cap G_w \Rightarrow (\|\kappa(w')\| \subseteq A_w \cap \cap G_w \ \& \ A_{w'} \cap \cap G_{w'} \subseteq A_w \cap \cap G_w)$.
- (ii) $w' \in (A_w \cap \cap G_w) - \cap G_{w'} \Rightarrow (v(\kappa(w')) = \emptyset \ \& \ A_{w'} \cap \cap G_{w'} = \|\beta(w')\|)$.
- (iii) $\forall w \ \exists v \ \forall w' \in A_w \cap \cap G_w \ \exists v' \in A_w \cap \cap G_w \ (vR_w v' \ \& \ w' \in A_{v'} \cap \cap G_{v'})$.

Proof. (i): Straightforward.

(ii): For the first part, suppose $w' \in v(\kappa(w'))$. Since $w' \in A_w \cap \cap G_w \Rightarrow v(\kappa(w')) \subseteq A_w \cap \cap G_w$, we have that $\|\kappa(w')\| \subseteq A_w \cap \cap G_w$. But then $w' \in \cap G_{w'}$. The second part follows immediately from the first part.

(iii): If $w \in A_w \cap \cap G_w$, then one may choose $v' = v = w$. If $w \notin A_w \cap \cap G_w$ and $v(\kappa(w)) = \emptyset$, then $A_w \cap \cap G_w = \|\beta(w)\| (\neq \emptyset)$ and one may choose $v' = v = w'$, where w' is any element of $\|\beta(w)\|$. Otherwise, one may choose $v = w'$, where w' is any element of $v(\kappa(w))$.

It remains to show that these lower bounds are also upper bounds. We do this - completely analogous to the case of justified true belief - by defining standard Kripke models for the systems such that out of each finite standard countermodel model a **JTB** countermodel can be constructed.

DEFINITION 10.2 Let R be a relation. R is *secondarily reflexive* iff $\forall x, y (xRy \Rightarrow yRy)$.
 R is *secondarily (reflexive or euclidean)* iff $\forall x, y (xRy \Rightarrow (yRy \vee \forall z, z' (yRz \wedge yRz' \Rightarrow zRz')))$.

DEFINITION 10.3 Let $M = \langle w_0, W, R, \models \rangle$ be a Kripke model.
 $P \subseteq PL$ is called *M-neutral* iff $w_0 R w \Rightarrow \exists w' w_0 R w' R w \wedge (w' \models p \Leftrightarrow p \in P)$.
 M is called *neutral* iff there exists an M -neutral $P \subseteq PL$.

DEFINITION 10.4 A **JB** model is a Kripke model which is serial, transitive, incestual, neutral and secondarily (reflexive or euclidean). A **JB'** model is a secondarily reflexive **JB** model.

PROPOSITION 10.5 **JB** (resp. **JB'**) is sound w.r.t. **JB** models (resp. **JB'** models).

Proof. Straightforward.

PROPOSITION 10.6 A canonical model for **JB** (resp. **JB'**) is a **JB** model (resp. **JB'** model).

Proof. Let us sketch the proof of the neutrality of the canonical models:

By (N), $\text{CON}(\{\varphi_0 \mid \diamond\varphi, \Box(\diamond\varphi \supset \varphi_0) \in w_0\} \cup \{\psi \mid \Box\psi \in w_0\})$. Let $P \subseteq \text{PL}$ be s.t. $\text{CON}(|P| \cup \{\varphi_0 \mid \diamond\varphi, \Box(\diamond\varphi \supset \varphi_0) \in w_0\} \cup \{\psi \mid \Box\psi \in w_0\})$, where $|P| = \{p \mid p \in P\} \cup \{\neg p \mid p \notin P\}$. Let $w_0 R w$, then $\text{CON}(|P| \cup \{\diamond\chi \mid \chi \in w\} \cup \{\psi \mid \Box\psi \in w_0\})$. Define w' to be the maximal consistent extension of $|P| \cup \{\diamond\chi \mid \chi \in w\} \cup \{\psi \mid \Box\psi \in w_0\}$. Then $w_0 R w' R w$ and $p \in w' \Leftrightarrow p \in P$.

COROLLARY 10.7

- (i) $\vdash_{\mathbf{JB}} \varphi \Leftrightarrow \models_{\mathbf{JB}} \varphi$
- (ii) $\vdash_{\mathbf{JB}'} \varphi \Leftrightarrow \models_{\mathbf{JB}'} \varphi$

Notice that neutrality is not a frame-property. In fact, **JB** and **JB'** are incomplete systems, since any frame of **JB** or **JB'** has to satisfy $\Box(\Box\varphi \vee \Box\psi) \supset (\Box\varphi \vee \Box\psi)$, which is not valid in **JB** models nor in **JB'** models.

LEMMA 10.8 **JB** and **JB'** have the finite model property.

Proof. Define wR^*w' iff $(w \models \Box\varphi \Rightarrow w' \models \varphi \wedge \Box\varphi)$ and, if $w \neq w_0$, $(w \models \Box\varphi \wedge \neg\varphi \Rightarrow (w' \models \diamond\psi \Rightarrow w' \models \Box\diamond\psi))$. (The second part is redundant in the case of **JB'**.)

Let us call **JTB** models which satisfy $\forall w \in W \cup(\kappa(w)) \neq \emptyset$ **JTB'** models.

LEMMA 10.9

- (i) \exists finite **JB** countermodel to $\varphi \in \mathcal{L}_{\mathbf{B}^j} \Rightarrow \exists$ **JTB** countermodel to φ .
- (ii) \exists finite **JB'** countermodel to $\varphi \in \mathcal{L}_{\mathbf{B}^j} \Rightarrow \exists$ **JTB'** countermodel to φ .

Proof. Slightly more complex than, but similar to lemma 8.2.

Hence we may conclude:

PROPOSITION 10.10

- (i) $\forall \varphi \in \mathcal{L}_{\mathbf{B}^j} (\models_{\mathbf{JB}} \varphi \Leftrightarrow \models_{\mathbf{JTB}} \varphi)$.
- (ii) $\forall \varphi \in \mathcal{L}_{\mathbf{B}^j} (\models_{\mathbf{JB}'} \varphi \Leftrightarrow \models_{\mathbf{JTB}'} \varphi)$.

Acknowledgments

I would like to thank Albert Visser for commenting on preliminary versions of this paper. Parts of the material in this paper have been presented at the (Dutch) National Working Group on Non-Monotonic Reasoning and at the European Workshop JELIA '90. The investigations were supported by the Foundation for Philosophical Research (SWON), which is subsidized by the Netherlands Organization for Scientific Research (NWO).

References

- Chellas, B. 1980. *Modal Logic: An Introduction*, Cambridge: Cambridge UP.
- Dretske, F. 1981. *Knowledge and the Flow of Information*, Oxford: Basic Blackwell.
- Gettier, E. 1963. Is justified true belief knowledge? *Analysis* 23:121-123.
- Halpern, J. and Moses, Y. 1984. Knowledge and common knowledge in a distributed environment. *Proc. 3rd ACM Conf. on Principles of Distributed Computing*. pp. 50-61.
- van der Hoek, W. 1991. Systems for knowledge and beliefs. *Logics in AI*, J. van Eijck (ed.). Berlin: Springer, pp. 267-281.
- Kraus, S. and Lehmann, D. 1986. Knowledge, belief and time. *Proc. 13th ICALP*, C. Krott (ed.). Berlin: Springer, pp. 186-195.
- Kreisel, G. 1967. Informal rigour and completeness proofs. *Problems in the Philosophy of Mathematics*, I. Lakatos (ed.). Amsterdam: North-Holland, pp. 138-171.
- Lehrer, K. 1990. *Theory of Knowledge*, London: Routledge.
- Lenzen, W. 1978. *Recent Work in Epistemic Logic*, Amsterdam: North-Holland.
- Lenzen, W. 1979. Epistemologische Betrachtungen zu [S4,S5]. *Erkenntnis* 14:33-56.
- Levesque, H. 1984. A logic of implicit and explicit belief. *Proc. AAAI-84*, Austin TX, pp. 198-202.
- Shoham, Y. and Moses, Y. 1989. Belief as defeasible knowledge. *Proc. IJCAI 89*, Los Altos: Morgan Kaufmann, pp. 1168-1173.
- Voorbraak, F. 1991. The logic of objective knowledge and rational belief. *Logics in AI*, J. van Eijck (ed.), Berlin: Springer, pp. 499-515.