# Epistemic Logic

Rasmus K. Rendsvig and John Symons

August 9, 2018

Epistemic logic is a subfield of epistemology concerned with logical approaches to knowledge, belief and related notions. Though any logic with an epistemic interpretation may be called an *epistemic logic*, the most widespread type of epistemic logics in use at present are modal logics. Knowledge and belief are represented via the modal operators $K$ and $B$, often with a subscript indicating the agent that holds the attitude. Formulas $K_a\varphi$ and $B_a\varphi$ are then read "agent $a$ knows that phi" and "agent $a$ believes that hi", respectively. Epistemic logic allows the formal exploration of the implications of epistemic principles. For example, the formula $K_a\varphi \to \varphi$ states that what is known is true, while $K_a\varphi \to K_aK_a\varphi$ states that what is known is known to be known. The semantics of epistemic logic are typically given in terms of possible worlds *via* Kripke models such that the formula $K_a\varphi$ is read to assert that $\varphi$ is true in all worlds agent $a$ considers epistemically possible relative to its current information. The central problems that have concerned epistemic logicians include, for example, determining which epistemic principles are most appropriate for characterizing knowledge and belief, the logical relations between different conceptions of knowledge and belief, and the epistemic features of groups of agents. Beyond philosophy proper, epistemic logic flourishes in theoretical computer science, economics, and related fields.

## Contents

# 1 Introduction

Aristotelian texts set the groundwork for discussions of the logic of knowledge and belief, particularly *De Sophisiticis Elenchis* as well as the *Prior* and *Posterior Analytics*. While Aristotle addressed the four alethic modes of possibility, necessity, impossibility, and contingency, Buridan, Pseudo Scotus, Ockham, and Ralph Strode, helped to extend Aristotle's insights to epistemic themes and problems [10, 33]. During this period, the Pseudo-Scot and William of Ockham supplemented Aristotle's study of mental acts of cognition and volition (see [10, 130]). Ivan Boh's studies of the history of fourteenth and fifteenth century investigations into epistemic logic provide a good account of the topic, especially his *Epistemic Logic in the Later Middle Ages* [10].

According to Boh, the English philosopher Ralph Strode formulated a fully general system of propositional epistemic rules in his influential 1387 book *Consequences* [10, 135]. Strode's presentation built on the earlier logical treatises of Ockham and Burley. Problems of epistemic logic were also discussed between the 1330s and 1360s by the so-called Oxford Calculators, most prominently by William Heytesbury and Richard Kilvington. By the fifteenth century, Paul of Venice and other Italian philosophers also engaged in sophisticated reflection on the relationship between knowledge, truth, and ontology.

Discussions of epistemic logic during the medieval period share a similar set of foundational assumptions with contemporary discussions. Most importantly, medieval philosophers explored the connection between knowledge and veracity: If I know $p$, then $p$ is true. Furthermore, many medieval discussions begin with an assumption similar to G.E. Moore's observation that an epistemic agent cannot coherently assert "$p$ but I do not believe (know) $p$". Sentences of this form are generally referred to as *Moore sentences*.

Modern treatments of the logic of knowledge and belief grew out of the work of philosophers and logicians writing from 1948 through the 1950s. Rudolf Carnap, Jerzy Łoś, Arthur Prior, Nicholas Rescher, G.H. von Wright and others recognized that our discourse concerning knowledge and belief admits of an axiomatic-deductive treatment. Among the many important papers that appeared in the 1950s, von Wright's seminal work (1951)[47] is widely acknowledged as having initiated the formal study of epistemic logic as we know it today. Von Wright's insights were extended by Jaakko Hintikka in his book *Knowledge*

*and Belief*: *An Introduction to the Logic of the Two Notions* (1962), [25]. Hintikka provided a way of interpreting epistemic concepts in terms of possible world semantics and as such it has served as the foundational text for the study of epistemic logic ever since.

In the 1980s and 1990s, epistemic logicians focused on the logical properties of systems containing groups of knowers and later still on the epistemic features of so-called "multi-modal" contexts. Since the 1990s work in dynamic epistemic logic has extended traditional epistemic logic by modeling the dynamic process of knowledge acquisition and belief revision. In the past two decades, epistemic logic has come to comprise a broad set of formal approaches to the interdisciplinary study of knowledge and belief.

Interest in epistemic logic extends well beyond philosophers. Recent decades have seen a great deal of interdisciplinary attention to epistemic logic with economists and computer scientists actively developing the field together with logicians and philosophers. In 1995 two important books signaled the fertile interplay between computer science and epistemic logic: Fagin, Halpern, Moses, and Vardi (1995) [16] and Meyer and van der Hoek (1995) [38]. Work by computer scientists has become increasingly central to epistemic logic in the intervening years.

Among philosophers, there is increased attention to the interplay between these formal approaches and traditional epistemological problems (See for example, [7], [23], [44], [30]).

Several introductory texts on epistemic logic exist, e.g. [14, 13, 17, 39].

## 2 The Modal Approach to Knowledge

Until relatively recently, epistemic logic focused almost exclusively on propositional knowledge. In cases of propositional knowledge, an agent or a group of agents bears the propositional attitude of knowing towards some proposition. For example, when one says: "Zoe knows that there is a hen in the yard", one asserts that Zoe is the agent who bears the propositional attitude *knowing* towards the proposition expressed by the English sentence "there is a hen in the yard". Now imagine that Zoe does not know whether there is a hen in the yard. For example, it might be the case that she has no access to information about whether there is or is not a hen in the yard. In this case her lack of information means that she will consider two scenarios as being possible, one in which there is a hen in the yard and one in which there is not.

Perhaps she has some practical decision that involves not only hens but also the presence of frightening dogs in the yard. She might wish to feed the hens but will only do so if there is no dog in the yard. If she were ignorant of whether there is a dog in the yard, the number of scenarios she must consider in her

deliberations grows to four. Clearly, one needs to consider epistemic alternatives when one does not have complete information concerning the situations that are relevant to one's decisions. As we shall see below, possible worlds semantics has provided a useful framework for understanding the manner in which agents can reason about epistemic alternatives.

While epistemic logicians had traditionally focused on *knowing that,* one finds a range of other uses of knowledge in natural language. As Wang [49] points out, the expressions *knowing how*, *knowing what*, *knowing why* are very common, appearing almost just as frequently (sometimes more frequently) in spoken and written language as *knowing that*. Recently non-standard epistemic logics of such expressions have been developed, though *knowing who* constructions are present in Hintikka's *Knowledge and Belief* [25] (see also [9, 41]). Thus, beyond propositional knowledge, epistemic logic also suggests ways to systematize the logic of questions and answers (Brendan knows why the dog barked). It also provides insight into the relationships between multiple modes of identification (Zoe knows that this man is the president). Here, the agent can be said to know a fact relating multiple modes of identification insofar as she correctly identifies the president, who she might know from stories in the newspaper with the man she sees standing in front of her, who she identifies as an object in her visual field [28]. Epistemic logic may also provide insight into questions of procedural "know-how" (Brendan knows how to change a fuse). For example, knowing how to $\varphi$ can be understood to be equivalent to the claim that there exists a way such that an agent knows that it is a way to ensure that $\varphi$ (See [49, 50]). Work concerning the justifications of knowledge have also been undertaken by combinations of *justification logic* with epistemic logic: See e.g. [2, 42].

There is ongoing work on these and other topics, and new developments are appearing steadily.

## 2.1 The Formal Language of Epistemic Logic

Recent work in epistemic logic relies on a modal conception of knowledge. In order to be clear about the role of modality in epistemic logic it is helpful to introduce the basic elements of the modern formalism. For the sake of simplicity we begin with the case of knowledge and belief for a single agent, postponing consideration of multiple agents to Section 3,

A prototypical epistemic logic language is given by first fixing a set of *propositional variables* $p_1, p_2, ....$ In applications of epistemic logic, propositional variables are given specific interpretations: For example, $p_1$ could be taken to represent the proposition "there is a hen in the yard" and $p_2$ the proposition "there is a dog in the yard", etc. The propositional variables represent propositions which are represented in no finer detail in the formal language. As such, they are therefore often referred to as *atomic propositions* or simply *atoms*. Let *Atom* denote the set of atomic propositions.

Apart from the atomic propositions, epistemic logic supplements the language of propositional logic with a modal operator, $K_a$, for knowledge and $B_a$,

for belief.

$K_a\varphi$ reads "Agent $a$ knows that $\varphi$"

and similarly

$B_a\varphi$ reads "Agent $a$ believes that $\varphi$."

In many recent publications on epistemic logic, the full set of formulas in the language is given using a so-called *Backus-Naur form*. This is simply a notational technique derived from computer science that provides a recursive definition of the formulas deemed grammatically 'correct', i.e., the set of *well-formed formulas*:

$$\varphi := p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid B_a\varphi, \text{ for } p \in Atom.$$

This says that $\varphi$ is $p$, if $p$ is an atom. $\neg\varphi$ is a well-formed formula if $\varphi$ is already a well-formed formula. The symbol '$\neg$' is a negation and '$\wedge$' a conjunction: $\neg\varphi$ reads 'not $\varphi$' while $(\varphi \wedge \psi)$ reads '$\varphi$ and $\psi$'. We will call this basic language that includes both a *K*nowledge and a *B*elief operator, $\mathcal{L}_{KB}$. As in propositional logic, additional connectives are defined from $\neg$ and $\wedge$: Typical notation is '$\vee$' for 'or', '$\rightarrow$' for 'if..., then ...' and '$\leftrightarrow$' for '... if, and only if, ...'. Also typically $\top$ ('top') and $\bot$ ('bottom') is used to denote the constantly true proposition and the constantly false proposition, respectively.

As we shall see below, $K_a\varphi$ is read as stating that $\varphi$ holds in *all* of the worlds accessible to $a$. In this sense, $K$ can be regarded as behaving similarly to the 'box' operator, $\Box$, often used to denote necessity. In evaluating $K_a\varphi$ at a possible world $w$, one is in effect evaluating a *universal quantification* over all the worlds accessible from $w$. The universal quantifier $\forall$ in first-order logic has the existential quantifier $\exists$ as its *dual*: This means that the quantifiers are mutually definable by taking either $\forall$ as primitive and defining $\exists x\varphi$ as short for $\neg\forall x\neg\varphi$ or by taking $\exists$ as primitive and defining $\forall x\varphi$ as $\neg\exists x\neg\varphi$. In the case of $K_a$, it may be seen that the formula $\neg K_a\neg\varphi$ makes an *existential quantification*: It says that there *exists* an accessible world that satisfies $\varphi$. In the literature, a dual operator for $K_a$ is often introduced. The typical notation for $\neg K_a\neg$ includes $\langle K_a\rangle$ and $\widehat{K}_a$. This notation mimics the diamond-shape $\Diamond$, which is the standard dual operator to the box $\Box$, which in turn is standard notation for the universally quantifying modal operator (see the entry on modal logic).

More expressive languages in epistemic logic involve the addition of operators for various notions of group knowledge (see Section 3). For example, as we discuss below, the *common knowledge* operator and so-called *dynamic* operators are important additions to the language of epistemic logic. Dynamic operators can indicate for example the *truthful public announcement* of $\varphi$: $[\varphi!]$. A formula $[\varphi!]\psi$ is read "if $\varphi$ is truthfully announced to everybody, then after the announcement, $\psi$ is the case." The question of what kinds of expressive power is added with the addition of operators is a research topic that is actively being investigated in dynamic epistemic logic. So, for example, adding $[\varphi!]$ by itself to $\mathcal{L}_{KB}$ does *not* add expressive power, but in a language that also includes common knowledge, it does.

## 2.2 Higher-Order Attitudes

Notice that for example $K_a K_a p$ is a formula in the language we introduced above. It states that agent $a$ knows that agent $a$ knows that $p$ is the case. Formula with *nested* epistemic operators of this kind express a *higher-order* attitude: an attitude concerning the the attitude of some agent.

Higher-order attitudes is a recurring theme in epistemic logic. The aforementioned Moore sentences—e.g. $B_a(p \wedge B_a \neg p)$—express a higher-order attitude. So do many of the epistemic principles discussed in the literature and below. Consider the following prominent epistemic principle involving higher-order knowledge: $K_a\varphi \rightarrow K_a K_a \varphi$. Is it reasonable to require that knowledge satisfies this scheme, i.e., that if somebody knows $\varphi$, then they know that they know $\varphi$? In part, we might hesitate before accepting this principle in virtue of the higher-order attitude involved. This is a matter of ongoing discussion in epistemic logic and epistemology.

## 2.3 The Partition Principle and Modal Semantics

The semantics of the formal language introduced above is generally presented in terms of so-called possible worlds. In epistemic logic possible worlds are interpreted as epistemic alternatives. Hintikka was the first to explicitly articulate such an approach (1962)[25]. This is another central feature of his approach to epistemology which continues to inform developments today. It may be stated, simplified[1], as follows:

> **Partition Principle:** Any propositional attitude partitions the set of possible worlds into those that are in accordance with the attitude those that are not.

The partition principle may be used to provide a semantics for the knowledge operator. Informally,

> $K_a\varphi$ is true in world $w$ if, and only if, $\varphi$ is true in every world $w'$ compatible with what $a$ knows at $w$.

Here, agent $a$ knows that $\varphi$ just in case the agent has information that rules out every possibility of error—rules out every case where $\neg\varphi$.

[1] In his 1969 [26, p. ], Hintikka writes "My basic assumption (slightly oversimplified) is that an attribution of any propositional attitude to the person in question involves a division of all possible worlds (more precisely, all the worlds which we can distinguish in the part of language we use in making the attribution) into two classes: into those possible worlds which are in accordance with the attitude in question and into those worlds which are incompatible with it." A similar view was held throughout his authorship, echoed in 2007 [24] by "What the concept of knowledge involves in a purely logical perspective is thus a dichotomy of the space of all possible scenarios into those that are compatible with what I know and those that are incompatible with my knowledge. This observation is all we need for most of epistemic logic."

## 2.4 Kripke Models and The Indistinguishability Interpretation of Knowledge

Since the 1960s *Kripke models*—defined below—have served as the basis of the most widely used semantics for all varieties of modal logic. The use of Kripke models in the representation of epistemic concepts involves taking a philosophical stance with respect to those concepts. One widespread interpretation, especially in theoretical economics and theoretical computer science, understands knowledge in terms of informational indistinguishability between possible worlds. What we will refer to here as the *indistinguishability interpretation* goes back at least to Lehmann (1984) [35].

As the indistinguishability interpretation concerns knowledge, but not belief, we will be working with a language without belief operators. Therefore, let the language $\mathcal{L}_K$ be given by the Backus-Naur form

$$\varphi := p \,|\, \neg\varphi \,|\, (\varphi \wedge \varphi) \,|\, K_a\varphi \text{ for } p \in \mathit{Atom}.$$

As we shall see, the indistinguishability interpretation involves very stringent requirements in order for something to qualify as knowledge. We introduce it here for pedagogical purposes, putting the formal details of the interpretation in place so as to introduce and explain relatively less extreme positions thereafter.

Consider again the case of Zoe, the hen and the dog. The example involves two propositions, which we will identify with the formal atoms:

$p$ read as "there is a hen in the yard."
and
$q$ read as "there is a dog in the yard."

It is worth emphasizing that for the purposes of our formalization of this scenario, these two are the *only* propositions of interest. We are restricting our attention to $\mathit{Atom} = \{p, q\}$. In early presentations of epistemic logic and in much of standard epistemic logic at present, *all* the atoms of interest are included from the outset. Obviously, this is an idealized scenario. It is important to notice what this approach leaves out. Considerations that are not captured in this way include the appearance of novel atoms; the idea that other atomic propositions might be introduced at some future state via some process of learning for example, or the question of an agent's awareness of propositions; the scenario in which an agent might be temporarily *unaware* of some atom due to some psychological or other factor (see Sec. 4 for references to so-called *awareness logic*). For now, the main point is that standard epistemic logic begins with the assumption that the set *Atom* exhausts the space of propositions for the agent.

With two atoms, there are four different ways a world could consistently be. We can depict each by a box:

The four boxes may be formally represented by a set $W = \{w_1, w_2, w_3, w_4\}$, typically called a set of ***possible worlds***. Each world is further labeled with the atoms true at that world. They are labeled by a function $V$, the ***valuation***. The valuation specifies which atoms are true at each world in the following way: Given an atom $p$, $V(p)$ is the subset of worlds at which $p$ is true.[2] That $w_1$ is labeled with $p$ and $q$ thus means that $w_1 \in V(p)$ and $w_1 \in V(q)$. In the illustration, $V(p) = \{w_1, w_2\}$ and $V(q) = \{w_1, w_3\}$.

For presentational purposes, assume that there really is a hen in the yard, but no dog. Then $w_2$ would represent the ***actual world*** of the model. In illustrations, the actual world is commonly highlighted:


Now, assume that the hen is always clucking, but that the dog never barks, and that although Zoe has acute hearing, she cannot see the yard. Then there are certain possible worlds that Zoe cannot *distinguish*: possible ways things may be which she cannot tell apart. For example, being in the world with only a hen $(p, \neg q)$, Zoe cannot tell if she is in the world with both hen and dog $(p, q)$: her situation is such that Zoe is aware of two ways things could be but her information does not allow her to eliminate either.

To illustrate that one possible world cannot be distinguished from another, an arrow is typically drawn from the former to the latter:


Here, arrows represent a *binary relation* on possible worlds. In modal logic in general, it is referred to as the ***accessibility relation***. Under the indistinguishability interpretation of epistemic logic, it is sometimes called the ***indistinguishability relation***. Formally, denote the relation $R_a$, with the subscript showing the relation belongs to agent $a$. The relation is a subset of the set of *ordered pairs* of possible worlds, $\{(w, w') \colon w, w' \in W\}$. One world $w$ "points" to another $w'$ if $(w, w') \in R_a$. In this case, $w'$ is said to be *accessible* (*indistinguishable*) from $w$. In the literature, this is often written $wR_aw'$ or $R_aww'$. The notation '$w' \in R_a(w)$' is also common: the set $R_a(w)$ is then the worlds accessible from $w$, i.e., $R_a(w) := \{w' \in W : (w, w') \in R_a\}$. A final note: the set $\{(w, w') \colon w, w' \in W\}$ is often written $W \times W$, the *Cartesian product* of $W$ with itself.

[2] In the literature, the valuation is sometimes defined the other way around, i.e., defined not by it assigning to each atom a set of worlds, but by it assigning to each world a set of atoms. In the example, we would thus look at $V(w_1)$—the valuation of the world—rather than $V(p)$—the valuation of the atom. $V(w_1)$ would be $\{p, q\}$. The two approaches are equivalent.

A second alternative—used e.g. in the ***[entry on modal logic]*** is to let the valuation be a function from the set of pairs of worlds and atoms $W \times Atom$ to the set truth values for *true* and *false*, $\{T, F\}$. Then $V(w_1, p) = T$ while $V(w_2, q) = F$. Again, this approach is equivalent to the one used here.

For $R_a$ to faithfully represent a relation of indistinguishability, what worlds should it relate? If Zoe was plunged in $w_1$ for example, could she tell that she is not in $w_2$? No: the relation of indistinguishability is *symmetric*—if one cannot tell $a$ from $b$, neither can one tell $b$ from $a$. That a relation is symmetric is typically drawn by omitting arrow-heads altogether or by putting them in both directions:


Which of the remaining worlds are indistinguishable? Given that the hen is always clucking, Zoe has information that allows her to distinguish $w_1$ and $w_2$ from $w_3$ and $w_4$—and *vice versa*, cf. symmetry. Hence, no arrows between these. The worlds $w_3$ and $w_4$ are indistinguishable. This brings us to the following representation:


Since no information will ever allow Zoe to distinguish something from itself, any possible world is thus related to itself—the indistinguishably relation is *reflexive*:


The standard interpretation of the Zoe example in terms of a possible worlds model is now complete. Before turning to a general presentation of the indistinguishability interpretation, let us look at what Zoe knows.

Recall the informal modal semantics of the knowledge operator from above:

> $K_a\varphi$ is true in world $w$ if, and only if, $\varphi$ is true in every world $w'$ compatible with the information $a$ has at $w$.

To approach a formal definition, take '$w \vDash \varphi$' to mean that $\varphi$ is true in world $w$. Thus we can, define truth of $K_a\varphi$ in $w$ by

> $w \vDash K_a\varphi$ iff $w' \vDash \varphi$ for all $w'$ such that $wR_aw'$.

This definition states that $a$ knows $\varphi$ in world $w$ if, and only if, $\varphi$ is the case in all the worlds $w'$ which $a$ cannot distinguish from $w$.

So, where does that leave Zoe? First off, the definition allows us to evaluate her knowledge in each of the worlds, but seeing as $w_2$ is the actual world, it is the world of interest. Here are some examples of what we can say about Zoe's knowledge in $w_2$:

1. $w_2 \vDash K_ap$. Zoe knows that the hen is in the yard as all the worlds indistinguishable from $w_2$—that would be $w_1$ and $w_2$—make $p$ true.

2. $w_2 \vDash \neg K_a q$. Zoe does not know that the dog is in the yard, as one of the indistinguishable worlds—in fact $w_2$ itself—makes $q$ false.

3. $w_2 \vDash K_a K_a p$. Zoe knows that she knows $p$ because *a)* $w_2 \vDash K_a p$ (cf. 1.) and *b)* $w_1 \vDash K_a p$.

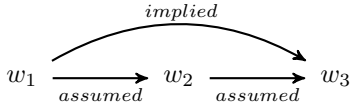
4. $w_2 \vDash K_a \neg K_a q$. Zoe knows that she does not know $q$ because *a)* $w_2 \vDash \neg K_a q$ (cf. 2.) and *b)* $w_1 \vDash \neg K_a q$.

We could say a lot more about Zoe's knowledge: every formula of the epistemic language without belief operators may be evaluated in the model. It thus represents all Zoe's higher-order information about her own knowledge—of which points 3. and 4. are the first examples.

One last ingredient is required before we can state the indistinguishability interpretation in its full generality. In the example above, it was shown that the indistinguishability relation was both *symmetric* and *reflexive*. Formally, these properties may be defined as follows:

> **Definition:** A binary relation $R \subseteq W \times W$ is
> *a)* ***reflexive*** iff for all $w \in W, wRw$,
> *b)* ***symmetric*** iff for all $w, w' \in W$, if $wRw'$, then $w'Rw$.

The missing ingredient is then the relational property of *transitivity*. 'Shorter than' is an example of a transitive property: Let $x$ be shorter than $y$, and let $y$ be shorter than $z$. Then $x$ must be shorter than $z$. So, given $w_1, w_2$ and $w_3$, if the relation $R$ holds between $w_1$ and $w_2$ and between $w_2$ and $w_3$, then the arrow between $w_1$ and $w_3$ is the consequence of requiring the relation to be transitive:




Formally, transitivity is defined as follows:

> **Definition:** A binary relation $R \subseteq W \times W$ is ***transitive*** iff for all $w, w', w'' \in W$, if $wRw'$ and $w'Rw''$, then $wRw''$

A relation that is both reflexive, symmetric and transitive is called an ***equivalence relation***.

With all the components in place, let us now define the Kripke model:

> **Definition:** A ***Kripke model*** for $\mathcal{L}_K$ is a tuple $M = (W, R, V)$ where
>
> - $W$ is a non-empty set of possible worlds,
> - $R$ is a binary relation on $W$, and
> - $V$: $Atom \longrightarrow \mathcal{P}(W)$ is a valuation.

In the definition, '$\mathcal{P}(W)$' denotes the *powerset* of $W$: It consists of all the subsets of $W$. Hence $V(p)$, the valuation of atom $p$ in the model $M$, is some subset of the possible worlds: Those where $p$ is true. In this general definition, $R$ can be any relation on $W$.

To specify which world is actual, one last parameter is added to the model. When the actual world is specified a Kripke model is commonly called *pointed*:

> **Definition:** A ***pointed Kripke model*** for $\mathcal{L}_K$ is a pair $(M, w)$ where
>
> - $M = (W, R, V)$ is a Kripke model, and
> - $w \in W$.

Finally, we may formally define the semantics that was somewhat loosely expressed above. This is done by defining a relation between pointed Kripke models and the formulas of the formal language. The relation is denoted '$\vDash$' and is often called the ***satisfaction relation***.

The definition then goes as follows:

> **Definition:** Let $M = (W, R_a, V)$ be a Kripke model for $\mathcal{L}_K$ and let $(M, w)$ be a pointed Kripke model. Then for all $p \in Atom$ and all $\varphi, \psi \in \mathcal{L}_K$
>
> $(M, w) \vDash p$ iff $w \in V(p)$
> $(M, w) \vDash \neg\varphi$ iff not $(M, w) \vDash \varphi$
> $(M, w) \vDash (\varphi \wedge \psi)$ iff $(M, w) \vDash \varphi$ and $(M, w) \vDash \psi$
> $(M, w) \vDash K_a\varphi$ iff $(M, w') \vDash \varphi$ for all $w' \in W$ such that $wR_aw'$.
>
> The formula $\varphi$ is ***satisfied*** in the pointed model $(M, w)$ iff $(M, w) \vDash \varphi$.

In full generality, the indistinguishability interpretation holds that for $K_a$ to capture knowledge, the relation $R_a$ must be an equivalence relation. A pointed Kripke model for which this is satisfied is often referred to as an ***epistemic state***. In epistemic states, the relation is denoted by a tilde with subscript: $\sim_a$.

Given pointed Kripke models and the indistinguishability interpretation, we have a semantic specification of one concept of knowledge. With this approach, we can build models of situations involving knowledge—as we did with the toy example of Zoe and the hens. We can use these models to determine what the agent does or does not know. We also have the formal foundations in place to begin asking questions concerning how the agent's knowledge or uncertainty develops when it receives *new information*, a topic studied in dynamic epistemic logic.

We may also ask more general questions concerning the concept of knowledge modeled using pointed Kripke models with indistinguishably relations: Instead of looking at a particular model at the time and asking which formulas the model makes true, we can ask which general principles all such models agree on.

## 2.5 Epistemological Principles in Epistemic Logic

Settling on the correct formal representation of knowledge involves reflecting carefully on the epistemological principles to which one is committed. An uncontroversial example of such a principle which most philosophers will accept is veridicality:

> *If a proposition is known, then it is true.*
> $K_a\varphi \to \varphi$.

In a formal context this principle can be understood to say that if $\varphi$ is known then it should always be satisfied in one's models. If it turns out that some of one's chosen models falsify the veridicality principle, then most philosophers would simply deem those models unacceptable.

Returning to pointed Kripke models, we can now ask which principles these models commit one to. In order to begin answering this question, we need to understand the most general features of our formalism. The strategy in modal logic in general (see [8]) is to abstract away from any given model's *contingent* features. Contingent features would include, for example, the specific number of worlds under consideration, the specific valuation of the atoms, and the choice of an actual world. In this case, the only features that are not contingent are those required by the general definition of a pointed Kripke model.

To abstract suitably, take a pointed Kripke model $(M, w) = (W, R, V, w)$. To determine whether the relation of this model is an equivalence relation we only need to consider the worlds and the relation. The pair of these elements constitute the fundamental level of the model and is called the *frame* of the model:

> **Definition:** Let $(M, w) = (W, R, V, w)$ be a pointed Kripke model. Then the pair $(W, R)$ is called the ***frame*** of $(M, w)$. Any model $(M', w')$ which shares the frame $(W, R)$ is said to be ***built on*** $(W, R)$.

Consider again the epistemic state for Zoe from above:


Several other models may be built on the same frame. The following are two examples:



With the notion of a frame, we may define the notion of validity of interest. It is the second term defined in the following:

**Definition:** A formula $\varphi$ is said to be ***valid in the frame*** $F = (W, R)$ iff every pointed Kripke model build on $F$ satisfies $\varphi$, i.e. iff for every $(M, w) = (F, V, w) = (W, R, V, w)$, $(M, w) \vDash \varphi$. A formula $\varphi$ is ***valid on the class of frames*** $\mathsf{F}$ (written $\mathsf{F} \vDash \varphi$) iff $\varphi$ is valid in every frame $F$ in $\mathsf{F}$.

The set of formulas valid on a class of frames $\mathsf{F}$ is called the ***logic*** of $\mathsf{F}$. Denote this logic—that is, the set $\{\varphi \in \mathcal{L}_K \colon \mathsf{F} \vDash \varphi\}$—by $\Lambda_{\mathsf{F}}$. This is a *semantic* approach to defining logics, each just a set of formulas. One may also define logics *proof-theoretically* by defining a logic as the set of formulas provable in some system. With logics as just sets of formulas, *soundness* and *completeness* results may then be expressed using set inclusion. To exemplify, let $\mathsf{A}$ be a set of axioms and write $\mathsf{A} \vdash \varphi$ when $\varphi$ is provable from $\mathsf{A}$ using some given set of deduction rules. Let the resulting logic—the set of theorems—be denoted $\Lambda_{\mathsf{A}}$. It is the set of formulas from $\mathcal{L}_K$ provable from $\mathsf{A}$—i.e., the set $\{\varphi \in \mathcal{L}_K \colon \mathsf{A} \vdash \varphi\}$. The logic $\Lambda_{\mathsf{A}}$ is sound with respect to $\mathsf{F}$ iff $\Lambda_{\mathsf{A}} \subseteq \Lambda_{\mathsf{F}}$ and complete with respect to $\mathsf{F}$ iff $\Lambda_{\mathsf{F}} \subseteq \Lambda_{\mathsf{A}}$.[3]

Returning to the indistinguishability interpretation of knowledge, we may then seek to find the epistemological principles which the interpretation is committed to. There is a trivial answer of little direct interest: Let $\mathsf{EQ}$ be the class of frames with equivalence relations. Then the logic of the indistinguishability interpretation is the set of formulas of $\mathcal{L}_K$ which are valid over $\mathsf{EQ}$—i.e., the set $\Lambda_{\mathsf{EQ}} := \{\varphi \in \mathcal{L}_K \colon \mathsf{EQ} \vDash \varphi\}$. Not very informative.

Taking an *axiomatic* approach to specifying the logic, however, yields a presentation in terms of easy to grasp principles. To start with the simplest, then the principle T states that knowledge is *factual*: If the agent knows $\varphi$, then $\varphi$ must be true. The more cumbersome K states that if the agent knows an implication, then if the agent knows the antecedent, it also knows the consequent. I.e., if we include the derivation rule *modus ponens* (from $\varphi \to \psi$ and $\varphi$, conclude $\psi$) as rule of our logic of knowledge, K states that knowledge is *closed under implication*. The principle B states that if $\varphi$ is true, then the agent knows that it considers $\varphi$ possible. Finally, 4 states that if the agent knows $\varphi$, then it knows that it knows $\varphi$. T, B and 4 in the table below (the names are historical and not all meaningful).

| | |
|---|---|
| K | $K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$ |
| T | $K_a\varphi \to \varphi$ |
| B | $\varphi \to K_a\widehat{K}_a\varphi$ |
| 4 | $K_a\varphi \to K_aK_a\varphi$ |

In lieu of epistemological intuitions, we could discuss a concept of knowledge by discussing these and other principles. Should we accept T as a principle that knowledge follows? What about the others? Before we proceed, let us first make clear how the four above principles relate to the indistinguishability

[3] This paragraph refers to *weak* completeness. For the difference between weak and *strong completeness*, and for general meta-theoretical results for modal logic, see e.g. [8].

interpretation. To do so, we need the notion of a *normal modal logic*. In the below definition, as in the above principles, we are technically using *formula schemas*. E.g., in $K_a\varphi \rightarrow \varphi$, the $\varphi$ is a variable ranging over formulas in $\mathcal{L}_K$. Thus, strictly speaking, $K_a\varphi \rightarrow \varphi$ is not a formula, but a *scheme* for obtaining a formula. A *modal instance* of $K_a\varphi \rightarrow \varphi$ is then the formula obtained by letting $\varphi$ be some concrete formula from $\mathcal{L}_K$. E.g., $K_ap \rightarrow p$ and $K_a(p \wedge K_aq) \rightarrow (p \wedge K_aq)$ are both modal instances of T.

> **Definition:** Let $\Lambda \subseteq \mathcal{L}_K$ be a set of modal formulas. Then $\Lambda$ is a ***normal modal logic*** iff $\Lambda$ satisfies all of the following:
>
> 1. $\Lambda$ contains all modal instances of the classical propositional tautologies.
> 2. $\Lambda$ contains all modal instances of K.
> 3. $\Lambda$ is closed under *modus ponens*: If $\varphi \in \Lambda$ and $\varphi \rightarrow \psi \in \Lambda$, then $\psi \in \Lambda$.
> 4. $\Lambda$ is closed under *generalization* (a.k.a. *necessitation*): If $\varphi \in \Lambda$, then $K_a\varphi \in \Lambda$.

There is a unique *smallest* normal modal logic (given the set *Atom*)—that which contains exactly what is required by the definition and *nothing more*. It is often called the ***minimal normal modal logic*** and is denoted by the boldface **K** (not to be confused with the non-boldface K denoting the schema).

The logic **K** is just a set of formulas from $\mathcal{L}_K$. I.e., $\mathbf{K} \subseteq \mathcal{L}_K$. Points 1.–4. gives a perspective on this set: They provide an *axiomatization*. Often, as below, the schema K is referred to as an axiom, though really the instantiations of K are axioms.

To **K**, we can add additional principles as axioms (axiom schemes) to obtain stronger logics (logics that have additional theorems: Logics $\Lambda$ for which $\mathbf{K} \subseteq \Lambda$). Of immediate interest is the logic called **S5**:

> **Definition:** The logic **S5** is the smallest normal modal logic containing all modal instances of T, B, and 4.

Here, then, is the relationship between the above four principles and the indistinguishability interpretation:

> **Theorem 1:** The logic **S5** is the logic of the class of pointed Kripke models build on frames with equivalence relations. I.e. $\mathbf{S5} = \Lambda_{\mathsf{EQ}}$.

What does this theorem tell us with respect to the principles of knowledge, then? In one direction it tells us that if one accepts the indistinguishably interpretation, then one has implicitly accepted the principles K, T, B and 4 as reasonable for knowledge. In the other direction, it tells us that if one finds that **S5** is the appropriate logic of knowledge *and* one finds that pointed Kripke models are the right way to semantically represent knowledge, then one must use an equivalence relation. Whether one should interpret this relation in terms of indistinguishability, though, is a matter on which logic is silent.

In discussing principles for knowledge, it may be that some of the four above seem acceptable, while others do not: One may disagree with the acceptability of B and 4, say, while accepting K and T. In understanding the relationship between **S5** and equivalence relations, a more fine-grained perspective is beneficial: Theorem 1 may be chopped into smaller pieces reflecting the contribution of the individual principles K, T, 4 and B to the equivalence requirement—i.e., that the relation should be at the same time reflexive, symmetric and transitive.

**Theorem 2:** Let $F = (W, R)$ be a frame. Then:

- All modal instances of K are valid in $F$.
- All modal instances of T are valid in $F$ iff $R$ is reflexive.
- All modal instances of B are valid in $F$ iff $R$ is symmetric.
- All modal instances of 4 are valid in $F$ iff $R$ is transitive.

There are a number of insights to gain from Theorem 2. First, if one wants to use *any* type of Kripke model to capture knowledge, then one must accept K. Skipping some details, one must in fact accept the full logic **K** as this is the logic of the class of *all* Kripke models (see e.g. [8]).

Second, the theorem shows that there is an intimate relationship between the individual epistemic principles and the properties on the relation. This, in turn, means that one, in general, may approach the 'logic' in epistemic logic from two sides—from intuitions about the accessibility relation or from intuitions about epistemic principles.

Several normal modal logical systems weaker than **S5** have been suggested in the literature. Here, we specify the logics by the set of their modal axioms. E.g., The logic **K** is given by $\{\mathrm{K}\}$, while **S5** is given by $\{\mathrm{K}, \mathrm{T}, \mathrm{B}, 4\}$. To establish nomenclature, the following table contains a selection of principles from the literature with the frame properties they characterize, cf. [3, 8], on the line below them. The frame conditions are not all straightforward.

In the table, the subscript on $R_a$ is omitted to ease readability, and so is the domain of quantification $W$ over which the worlds variables $x, y, z$ range.

| | |
|---|---|
| K | $K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$ |
| | None: *Not applicable* |
| D | $K_a\varphi \to \widehat{K}_a\varphi$ |
| | Serial: $\forall x \exists y, xRy$. |
| T | $K_a\varphi \to \varphi$ |
| | Reflexive: $\forall x, xRx$. |
| 4 | $K_a\varphi \to K_aK_a\varphi$ |
| | Transitive: $\forall x, y, z,$ if $xRy$ and $yRz$, then $xRz$. |
| B | $\varphi \to K_a\widehat{K}_a\varphi$ |
| | Symmetric: $\forall x, y,$ if $xRy$, then $yRx$. |
| 5 | $\neg K_a\varphi \to K_a\neg K_a\varphi$ |
| | Euclidean: $\forall x, y, z,$ if $xR_ay$ and $xR_az$, then $yRz$. |
| .2 | $\widehat{K}_aK_a\varphi \to K_a\widehat{K}_a\varphi$ |
| | Confluent: $\forall x, y,$ if $xRy$ and $xRy'$, then $\exists z, yRz$ and $y'Rz$. |
| .3 | $(\widehat{K}_a\varphi \wedge \widehat{K}_a\psi) \to (\widehat{K}_a(\varphi \wedge \widehat{K}_a\psi) \vee \widehat{K}_a(\varphi \wedge \psi) \vee \widehat{K}_a(\psi \wedge \widehat{K}_a\varphi))$ |
| | No branching to the right: $\forall x, y, z,$ if $xRy$ and $xRz$, then $yRz$ or $y = z$ or $zRy$ |
| .3.2 | $(\widehat{K}_a\varphi \wedge \widehat{K}_aK_a\psi) \to K_a(\widehat{K}_a\varphi \vee \psi)$ |
| | Semi-Euclidean: $\forall x, y, z,$ if $xRy$ and $xRz$, then $zRx$ or $yRz$. |
| .4 | $(\varphi \wedge \widehat{K}_aK_a\varphi) \to K_a\varphi$ |
| | Unknown to authors: *Not applicable* |

Table 1: Epistemic principles and their frame conditions.

Adding epistemic principles as axioms to the basic minimal normal modal logic **K** yields new, normal modal logics. A selection is:

| Logic names and axioms | |
|---|---|
| **K** | {K} |
| **T** | {K, T} |
| **D** | {K, D} |
| **KD4** | {K, D, 4} |
| **KD45** | {K, D, 4, 5} |
| **S4** | {K, T, 4} |
| **S4.2** | {K, T, 4, .2} |
| **S4.3** | {K, T, 4, .3} |
| **S4.4** | {K, T, 4, .4} |
| **S5** | {K, T, 5} |

Different axiomatic specifications may produce the same logic. Notice e.g. that the table's axiomatic specification $\{\mathrm{K}, \mathrm{T}, 5\}$ of **S5** does not match that given in the definition preceding Theorem 1, $\{\mathrm{K}, \mathrm{T}, \mathrm{B}, 4\}$. Note also, there is more than one axiomatization of **S5**: the axioms $\{\mathrm{K}, \mathrm{T}, 5\}$, $\{\mathrm{K}, \mathrm{T}, \mathrm{B}, 4\}$, $\{\mathrm{K}, \mathrm{D}, \mathrm{B}, 4\}$ and $\{\mathrm{K}, \mathrm{D}, \mathrm{B}, 5\}$ all give the **S5** logic, cf. e.g. [11]. An often seen variant is $\{\mathrm{K}, \mathrm{T}, 4, 5\}$. However, it is redundant to add it as all its instances can be proven from K, T and 5. But as both 4 and 5 capture important epistemic principles (see Sec. 2.6), 4 is often sometimes included for the sake of philosophical transparency. For more equivalences between modal logics, see e.g. the entry on modal logic or [11] or [8].

Logics may be stronger or weaker than each other, and knowing the frame properties of their axioms may help us to understand their relationship. For example, as 4 is derivable from $\{\mathrm{K}, \mathrm{T}, 5\}$, all the theorems of **S4** are derivable in **S5**. **S5** is thus *at least as strong* as **S4**. In fact, **S5** is also *strictly stronger*: It can prove things which **S4** cannot.

That **S5** may be axiomatized both by $\{\mathrm{K}, \mathrm{T}, \mathrm{B}, 4\}$ and $\{\mathrm{K}, \mathrm{T}, 5\}$ may be seen through the frame properties of the axioms: every reflexive and euclidean relation (T and 5) is an equivalence relation (T,B and 4). This also shows the redundancy of 4: If one has assumed a relation reflexive and euclidean, then it adds nothing new to additionally assume it to be transitive. In general, having an understanding of the interplay between relational properties is of great aid in seeing relationships between modal logics. For example, noticing that every reflexive relation is also serial means that all formulas valid on the class of serial models are also valid on the class of reflexive models. Hence, every theorem of **D** is thus a theorem of **T**. Hence **T** is at least as strong as **D** (i.e., $\mathbf{D} \subseteq \mathbf{T}$). That **T** is also strictly stronger (not $\mathbf{T} \subseteq \mathbf{D}$) can be shown by finding a serial, non-reflexive model which does not satisfy some theorem of **T** (for example $K_a p \to p$).

## 2.6 Principles of Knowledge and Belief

With the formal background of epistemic logic in place, it is straightforward to slightly vary the framework in order to accommodate the concept of belief. Return to the language $\mathcal{L}_{KB}$ of both knowledge and belief:

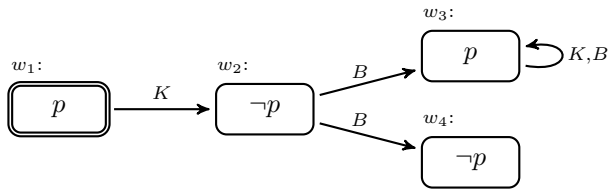
$$\varphi := p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\psi \mid B_a\psi, \text{ for } p \in \mathit{Atom}.$$

To interpret knowledge and belief formulas together in pointed Kripke models, all that is needed is an additional relation between possible worlds:

**Definition:** A ***pointed Kripke model*** for $\mathcal{L}_{KB}$ is a tuple $(M, w) = (W, R_K, R_B, V, w)$ where

- $W$ is a non-empty set of possible worlds,
- $R_K$ and $R_B$ are a binary relations on $W$,
- $V\colon \mathit{Atom} \longrightarrow \mathcal{P}(W)$ is a valuation, and
- $w \in W$.

$R_K$ is the relation for the knowledge operator and $R_B$ the relation for the belief operator. The definition makes no further assumptions about their properties. In the figure below we provide an illustration, where the arrows are labeled in accordance with the relation they correspond to. The reflexive loop at $w_3$ is a label indicating that it belongs to both relations, i.e., $(w_3, w_3) \in R_K$ and $(w_3, w_3) \in R_B$.




The satisfaction relation is defined as above, but with the obvious changes for knowledge and belief:

$(M, w) \vDash K_a\varphi$ iff $(M, w') \vDash \varphi$ for all $w' \in W$ such that $wR_Kw'$.
$(M, w) \vDash B_a\varphi$ iff $(M, w') \vDash \varphi$ for all $w' \in W$ such that $wR_Bw'$.

The indistinguishability interpretation puts very strong requirements on the accessibility relation for knowledge. These have now been stripped away and so has any commitment to the principles T, B, D, 4 and 5. Taking Kripke models as basic semantics, we are still committed to K, though this principle is not unproblematic as we shall see below in our discussion of the problem of logical omniscience.

Of the principles from Table 1, T, D, B, 4 and 5 have been discussed most extensively in the literature on epistemic logic, both as principles for knowledge and as principles for belief. The principle T for knowledge

$$K_a\varphi \rightarrow \varphi$$

is broadly accepted. Knowledge is commonly taken to be *veridical*—only true proposition can be known. For e.g. Hintikka [25] and Fagin et al. [16], the failure of T for belief is the defining difference between the two notions.

Though belief is not commonly taken to be veridical, believes are typically taken to be *consistent*. I.e., agents are taken to *never* believe the contradiction—that is, any formula equivalent with $(p \wedge \neg p)$—or $\bot$, for short. That believes should be consistent is then captured by the principle

$$\neg B_a\bot.$$

The principle $\neg B_a\bot$ is, on Kripke models, equivalent with the principle D, $B_a\varphi \rightarrow \widehat{B}_a\varphi$. Hence the validity of $\neg B_a\bot$ requires serial frames. Witness e.g. its failure in $w_1$ above: As there are no worlds accessible through $R_B$, *all* accessible worlds satisfy $\bot$. Hence $w_1$ satisfies $B_a\bot$, violating consistency. Notice also that $\neg B_a\bot$ may be re-written to $\widehat{B}_a\top$, which is true at a world just in case some world is accessible through $R_B$. Its validity thus ensures seriality.

Notice that the veridicality of knowledge ensures its consistency: Any reflexive frame is automatically serial. Hence accepting $K_a\varphi \to \varphi$ implies accepting $\neg K_a \bot$.

Of the principles D, 4 and 5, the two latter have received far the most attention, both for knowledge and for belief. They are commonly interpreted as governing of *principled access* to own mental states. The 4 principles

$$K_a\varphi \to K_aK_a\varphi$$
$$B_a\varphi \to B_aB_a\varphi$$

are often referred to as *principles of positive introspection*, or for knowledge the *'KK' principle*. Both principles are deemed acceptable by e.g. Hintikka [25] on grounds *different* from introspection. He argues based on an autoepistemic analysis of knowledge, using a non-Kripkean possible worlds semantics called *model systems*. Hintikka holds that when an agent commits to knowing $\varphi$, the agent commits to holding the same attitude no matter what new information the agent will encounter in future. This entails that in all the agent's epistemic alternatives—for Hintikka, all the model sets (partial descriptions of possible worlds) where the agent knows at least as much they now do—the agent still knows $\varphi$. As $K_a\varphi$ thus holds in all the agent's epistemic alternatives, Hintikka concludes that $K_aK_a\varphi$. Likewise Hintikka endorses 4 for belief, but Lenzen raises objections [36, Chap. 4].

Williamson argues against the general acceptability of the principle [51, Chap. 5] for a concept of knowledge based on slightly inexact observations, a so-called *margin of error principle*; see e.g. [3] for a short summary.

The 5 principles

$$\neg K_a\varphi \to K_a\neg K_a\varphi$$
$$\neg B_a\varphi \to B_a\neg B_a\varphi$$

are often referred to as *principles of negative introspection*. Negative introspection is quite controversial as it poses very high demands on knowledge and belief. The schema 5 may be seen as a *closed world assumption* [21]: The agent has complete overview of all the possible worlds and own information. If $\neg\psi$ is considered possible ($\widehat{K}_a\neg\psi$, i.e., $\neg K_a\psi$), then the agent knows it is considered possible ($K_a\neg K_a\psi$). Such a closed world assumption is natural when constructing hyper-rational agents in e.g. computer science or game theory, where the agents are assumed to reason as hard as logically possible about their own information when making decisions.

Arguing against 5 is Hintikka [25], using his conception of epistemic alternatives. Having accepted T for knowledge, 5 stands or falls with the assumption of a symmetric accessibility relation. But, Hintikka argus, the accessibility relation is not symmetric: If the agent possess some amount of information at model set $s_1$, then the model set $s_2$ where the agent has learned something more will be an epistemic alternative to $s_1$. But $s_1$ will not be an epistemic alternative to $s_2$, because in $s_1$, the agent does—by hypothesis—not know as much as it does in $s_2$. Hence the relation is not symmetric, so 5 is not a principle of knowledge, on Hintikka's account.

Given Hintikka's non-standard semantics, it is a bit difficult to pin down whether he would accept a normal modal logic as the logics of knowledge and belief, but if so, then **S4** and **KD4** would be the closest candidates (see [22] for this point). By contrast, for knowledge von Kutschera argued for **S4.4** (1976)[46], Lenzen suggested **S4.2** [36], van der Hoek argued for **S4.3** (1993) [1], and Fagin, Halpern, Moses and Vardi [16] and many others use **S5** for knowlegde and **KD45** for belief.

Beyond principles governing knowledge and principles governing belief, one may also consider principles governing the interplay between knowledge and belief. Three principles of interest are

KB1 $K_a\varphi \to B_a\varphi$
KB2 $B_a\varphi \to K_aB_a\varphi$
KB3 $B_a\varphi \to B_aK_a\varphi$

The principles KB1 and KB2 were introduced by Hintikka, who endorses both [25] noting that Plato is also committed to KB1 in *Theatetus*. The first principle, KB1, captures the intuition that knowledge is a stronger notion than belief. The second—like 4 and 5—captures the idea that one has privileged access to one's own beliefs. The third, stemming from Lenzen [36], captures the notion that beliefs are held with some kind of conviction: if something is believed, it is believed to be known.

Though the interaction principles KB1–KB3 may look innocent on their own, they may lead to counterintuitive conclusions when combined with specific logics of knowledge and belief. First, Voorbraak [48] shows that combining 5 for knowledge and D for belief with KB1, implies that

$$B_aK_a\varphi \to K_a\varphi$$

is a theorem of the resulting logic. Assuming that knowledge is truthful, this theorem entails that agents cannot believe to know something which happens to be false.

If additionally KB3 is added, the notions of knowledge and belief *collapse*. I.e., it may be proven that $B_a\varphi \to K_a\varphi$, which, in combination with KB1 entails that

$$B_a\varphi \leftrightarrow K_a\varphi.$$

Hence, the two notions have collapsed to one. This was stated in 1986, by Kraus and Lehmann [34].

If one is not interested in knowledge and belief collapsing, one must thus give something up: One cannot have both 5 for knowledge, D for belief and KB1 and KB3 governing their interaction. Again, results concerning correspondence between principles and relation properties may assist: In 1993, van der Hoek [1] showed based on a semantic analysis that where the four principles are jointly sufficient for collapse, *no subset of them is, too*. Giving up any one principle will thus eliminate the collapse. Weaking KB1 to hold only for non-modal formulas is also sufficient to avoid collapse, cf. [19].

For more on epistemic interaction principles, the principles .2, .3, .3.2. and .4, and relations to so-called *conditional beliefs*, see [3]. For an introduction to conditional beliefs and relations to several other types of knowledge from the philosophical literature, see [6]. The latter also includes discussion concerning the interdefinability of various notions, as does [18] for knowledge and (non-conditional) belief.

# 3 Knowledge in Groups

We human beings are preoccupied with the epistemic states of other agents. In ordinary life, we reason with varying degrees of success about what others know. We are especially concerned with what others know about us, and often specifically about what they know about what we know.

Does she know that I know where she buried the treasure?

Does she know that I know that she knows?

And so on.

Epistemic logic can reveal interesting epistemic features of systems involving groups of agents. In some cases, for example, emergent social phenomena depend on agents reasoning in particular ways about the knowledge and beliefs of other agents. As we have seen, traditional systems of epistemic logic applied only to single-agent cases. However, they can be extended to groups or multi-agent systems in a relatively straightforward manner.

As David Lewis noted in his book *Convention* [37] many prominent features of social life depend on agents assuming that the rules of some practice are matters of *common knowledge*. For example, drivers know that a red traffic light indicates that they should stop at an intersection. However, for the convention of traffic lights to be in place at all, it is first necessary that drivers must also know that other drivers know that *red* means *stop*. In addition, drivers must also know that everyone knows that everyone knows that .... The conventional role of traffic lights relies on all drivers knowing that all drivers know the rule, that the rule is a piece of common knowledge.

A variety of norms, social and linguistic practices, agent interactions and games presuppose common knowledge, first formalized by Aumann [4] and with earliest epistemic logical treatments by Lehmann [35] and by Halpern and Moses [20]. In order to see how epistemic logic sheds light on these phenomena, it is necessary to introduce a little more formalism. Following the standard treatment (see e.g. [16]), we can syntactically augment the language of propositional logic with $n$ knowledge operators, one for each agent involved in the group of agents under consideration. The primary difference between the semantics given

for a mono-agent and a multi-agent semantics is roughly that $n$ accessibility relations are introduced. A modal system for $n$ agents is obtained by joining together $n$ modal logics where for simplicity it may be assumed that the agents are homogenous in the sense that they may all be described by the same logical system. An epistemic logic for $n$ agents consists of $n$ copies of a certain modal logic. In such an extended epistemic logic it is possible to express that some agent in the group knows a certain fact that an agent knows that another agent knows a fact etc. It is possible to develop the logic even further: Not only may an agent know that another agent knows a fact, but they may all know this fact simultaneously.

## 3.1 Multi-Agent Languages and Models

To represent knowledge for a set $\mathcal{A}$ of $n$ agents, first let's stipulate a language. Let $\mathcal{L}_{Kn}$ be given by the *Backus-Naur form*

$$\varphi := p \,|\, \neg\varphi \,|\, (\varphi \wedge \varphi) \,|\, K_i\psi \text{ for } p \in Atom, i \in \mathcal{A}.$$

To represent knowledge for all $n$ agents jointly in pointed Kripke models, all that is needed is to add suitably many relations:

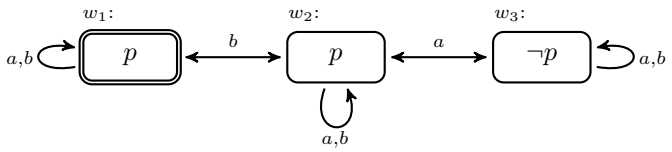
**Definition:** A ***pointed Kripke model*** for $\mathcal{L}_{Kn}$ is a tuple $(M, w) = (W, \{R_i\}_{i\in\mathcal{A}}, V, w)$ where

- $W$ is a non-empty set of possible worlds,
- For every $i \in \mathcal{A}$, $R_i$ is a binary relation on $W$,
- $V$: $Atom \longrightarrow \mathcal{P}(W)$ is a valuation, and
- $w \in W$.

To also incorporate beliefs, simply apply the same move as in the single agent case: augment the language and let there be two relations for each agent.

The definition uses a family of relations $\{R_i\}_{i\in\mathcal{A}}$. In the literature, the same is denoted $(W, R_i, V, w)_{i\in\mathcal{A}}$. Alternatively, $R$ is taken to be a function sending agents to relations, i.e., $R : \mathcal{A} \to \mathcal{P}(W \times W)$. Then for each $i \in \mathcal{A}$, $R(i)$ is a relation on $W$, often denoted $R_i$. These are stylistic choices.

When considering only a single agent, it is typically not relevant to include more worlds in $W$ than there are possible valuations of atoms. In multi-agent cases, this is not the case: to express the different forms of available higher-order knowledge, many copies of "the same" world are needed. Let us exemplify for $\mathcal{A} = \{a, b\}$, $Atom = \{p\}$ and each $R_i, i \in \mathcal{A}$, an equivalence relation. Let us represent that both $a$ and $b$ know $p$, but $b$ does not know that $a$ knows $p$ , i.e., $K_a p \wedge K_b p \wedge \neg K_b K_a p$. Then we need three worlds:



If we try to let $w_1$ play the role of $w_2$, then $a$ would lose knowledge in $p$: both $p$ worlds are needed. In general, if $W$ is assumed to have any fixed, finite size, there will be some higher-order information formula that cannot be satisfied in it.

## 3.2 Notions of Group Knowledge

Multi-agent systems are interesting for other reasons than to represent higher-order information. The individual agents' information may also be pooled to capture what the agents know jointly, as group knowledge (see [5] for a recent discussion). A standard notion is this style is *distributed knowledge*: The knowledge the group *would have* if the agents share all their individual knowledge. To represent it, augment the language $\mathcal{L}_{Kn}$ with operators

$$D_G \text{ for } G \subseteq \mathcal{A},$$

to make $D_G\varphi$ a well-formed formula. Where $G \subseteq \mathcal{A}$ is a group of agents, the formula $D_G\varphi$ reads that it is *distributed knowledge in the group $G$ that $\varphi$*.

To evaluate $D_G\varphi$, we define a new relation from those already present in the model. The idea behind the definition is that if some one agent has eliminated a world as an epistemic alternative, then so will the group. Define the relation as the intersection of the individual agents' relations:

$$R_G^D = \bigcap_{i \in G} R_i$$

In the three state model, $R_G^D$ contains only the three loops. To evaluate a distributed knowledge formula, use the same form as for other modal operators:

$$(M, w) \vDash D_G\varphi \text{ iff } (M, w') \vDash \varphi \text{ for all } w' \in W \text{ such that } wR_G^Dw'.$$

It may be the case that some very knowing agent knows all that is distributed knowledge in $G$, but it is not guaranteed. To capture that all the agents know $\varphi$, we could use the conjunction of the formulas $K_i\varphi$ for $\in \mathcal{A}$, i.e., $\bigwedge_{i\in\mathcal{A}} K_i\varphi$. This is a well-defined formula if $\mathcal{A}$ is finite (which it typically is). If $\mathcal{A}$ is not finite, then $\bigwedge_{i\in\mathcal{A}} K_i\varphi$ is not a formula in $\mathcal{L}_{Kn}$, as it has only finite conjunctions. As a shorthand for $\bigwedge_{i\in\mathcal{A}} K_i\varphi$, it is standard to introduce the *everybody knows* operator, $E_G$:

$$E_G\varphi := \bigwedge_{i\in\mathcal{A}} K_i\varphi.$$

In the three world model, $K_ap \wedge K_bp$, so $E_{\{a,b\}}p$.

That everybody knows something does not mean that this knowledge is shared between the members of the group. The three world model exemplifies this: Though $E_{\{a,b\}}p$, it also the case that $\neg K_bE_{\{a,b\}}p$.

To capture that there is no uncertainty in the group about $\varphi$ nor *any higher-order* uncertainty about $\varphi$ being known by all agents, no formula in the language $\mathcal{L}_{Kn}$ is enough. Consider the formula

$$E_G^k \varphi$$

where $E_G^k$ is short for $k$ iterations of the $E_G$ operator. Then for no natural number $k$ will the formula $E_G^k \varphi$ be enough: it could be the case that $b$ doesn't know it! To rectify this situation, one could try

$$\bigwedge_{k \in \mathbb{N}} E_G^k \varphi$$

but this is not a formula as $\mathcal{L}_{Kn}$ only contains finite conjunctions.

Hence, though the $E_G$ operator is definable in the language $\mathcal{L}_{Kn}$, a suitable notion of *common knowledge* is not. For that, we again need to define a new relation on our model. This time, we are interested in capturing that nobody considers $\varphi$ epistemically possible *anywhere*. To build the relation, we therefore first take the union the relations of all the agents in $G$, but this is not quite enough: to use the standard modal semantic clause, we must also be able to reach all of the worlds in this relation *in a single step*. Hence, let

$$R_G^C := \left( \bigcup_{i \in G} R_i \right)^*$$

where $(\cdot)^*$ is the operation of taking the *transitive closure*. If $R$ is a relation, then $(R)^*$ is $R$ plus all the pairs missing to make $R$ a transitive relation. Consider the three world model: With the relation $\bigcup_{i \in \{a,b\}} R_i$, we can reach $w_3$ from $w_1$ in two steps, stopping over at $w_2$. With $(\bigcup_{i \in \{a,b\}} R_i)^*$, $w_3$ is reachable in one step: By the newly added transitive link from $w_1$ to $w_3$.

To represent common knowledge, augment the *Backus-Naur form* of $\mathcal{L}_{Kn}$ with operators

$$C_G \text{ for } G \subseteq \mathcal{A},$$

to make $C_G \varphi$ a well-formed formula. Evaluate such formulas by the semantic clause

$$(M, w) \vDash C_G \varphi \text{ iff } (M, w') \vDash \varphi \text{ for all } w' \in W \text{ such that } w R_G^C w'.$$

Varying the properties of the accessibility relations $R_1, R_2, ..., R_n$, as described above results in different epistemic logics. For instance system **K** with common knowledge is determined by all frames, while system **S4** with common knowledge is determined by all reflexive and transitive frames. Similar results can be obtained for the remaining epistemic logics [16]. For more, consult the entry on common knowledge.

# 4 Logical Omniscience

The principal complaint against the approach taken by epistemic logicians is that it is committed to an excessively idealized picture of human reasoning. Critics have worried that the relational semantics of epistemic logic commits one to a closure property for an agent's knowledge that is implausibly strong given actual human reasoning abilities. The closure properties give rise to what has come to be called the problem of logical omniscience:

> Whenever an agent $c$ knows all of the formulas in a set $\Gamma$ and $A$ follows logically from $\Gamma$, then $c$ also knows $A$.

In particular, $c$ knows all theorems (letting $\Gamma = \emptyset$), and knows all the logical consequences of any formula that the agent knows (letting $\Gamma$ consist of a single formula). The concern here is that finite agents are constrained by limits on their cognitive capacities and reasoning abilities. The account of knowledge and belief that epistemic logic seems committed to involves superhuman abilities like knowing all the tautologies. Thus, the concern is that epistemic logic is simply unsuited to capturing actual knowledge and belief as these notions figure in ordinary human life.

Hintikka recognized a discrepancy between the rules of epistemic logic and the way the verb "to know" is ordinarily used already in the early pages of *Knowledge and Belief*. He pointed out that "it is clearly inadmissible to infer "he knows that $q$" from "he knows that $p$" solely on the basis that $q$ follows logically from $p$, for the person in question may fail to see that $p$ entails $q$, particularly if $p$ and $q$ are relatively complicated statements" (1962, 30-1). Hintikka's first reaction to what came to be called the problem of logical omniscience was to see the discrepancy between ordinary usage of terms like 'consistency' and formal treatments of knowledge as indicating a problem with our ordinary terminology. If a person knows the axioms of a mathematical theory but is unable to state the distant consequences of the theory, Hintikka denied that it is appropriate to call that person inconsistent. In ordinary human affairs, Hintikka claimed, the charge of inconsistency when directed towards an agent has the connotation of being irrational or dishonest. Thus, from Hintikka's perspective we should choose some other term to capture the situation of someone who is rational and amenable to persuasion or correction but not logically omniscient. Non-omniscient, rational agents can be in a position to say that "I know that $p$ but I don't know whether $q$" even in case $q$ can be shown to be entailed logically by $p$. He then suggests that $q$ should be regarded as *defensible* given the agent's knowledge and the denial of $q$ should be regarded as *indefensible*. This choice of terminology was criticized insofar as it attaches the pejorative *indefensible* to some set of proposition, even though the fault actually lies in the agent's cognitive capacities ([12, 29, 31]).

Hintikka's early epistemic logic can be understood as a way of reasoning about what is implicit in an agent's knowledge even in cases where the agent itself is unable to determine what is implicit. Such an approach risks being excessively idealized and its relevance for understanding human epistemic circumstances can be challenged on these grounds.

Few philosophers were satisfied with Hintikka's attempt to revise our ordinary use of the term 'consistent' as he presented it in *Knowledge and Belief*. However, he and others soon provided more popular ways of dealing with logical omniscience. In the 1970s responses to the problem of logical omniscience introduced semantical entities that explain why the agent appears to be, but in fact is not really guilty of logical omniscience. Hintikka called these entities 'impossible possible worlds' [27] (see also the entry on impossible worlds and [32]). The basic idea is that an agent may mistakenly count among the worlds consistent with its knowledge, some worlds containing logical contradictions. The mistake is simply a product of the agent's limited resources; the agent may not be in a position to detect the contradiction and may erroneously count them as genuine possibilities. In some respects, this approach can be understood as an extension of the aforementioned response to logical omniscience that Hintikka had already outlined in *Knowledge and Belief*.

In the same spirit, entities called 'seemingly possible' worlds are introduced by Rantala [40] in his urn-model analysis of logical omniscience. Allowing impossible possible worlds or seemingly possible worlds in which the semantic valuation of the formulas is arbitrary to a certain extent provides a way of making the appearance of logical omniscience less threatening. After all, on any realistic account of epistemic agency, the agent is likely to consider (albeit inadvertantly) worlds in which the laws of logic do not hold. Since no real epistemic principles hold broadly enough to encompass impossible and seemingly possible worlds, some conditions must be applied to epistemic models such that they cohere with epistemic principles (for criticism of this approach see [31, 336-7]).

Alternatively to designing logics in which the knowledge operators do not exhibit logical omniscience, *awareness logic* offers an alternative: Change the interpretation of $K_a\varphi$ from "$a$ knows that $\varphi$" to "$a$ *implicitely* knows that $\varphi$" and take *explicit* knowledge that $\varphi$ to be implicit knowledge that $\varphi$ *and* awareness of $\varphi$. With awareness not closed under logical consequence, such a move allows for notion of explicit knowledge not logically omniscient. As agents neither have to compute their implicit knowledge nor can they be held responsible for answering queries based on it, logical omniscience is problematic only for explicit knowledge, the *problem* of logical omniscience is thus averted. Though logical omniscience is an epistemological condition for implicit knowledge, the agent itself may actually fail to realize this condition. For more on awareness logic, see e.g. the seminal [15] or [45, 43] for overviews.

Debates about the various kinds of idealization involved in epistemic logic are ongoing in both philosophical and interdisciplinary contexts.

## References


[1] Wiebe Systems for knowledge and belief. *Journal of Logic and Computation*, 3:173–195, 1993. 2.6

[2] Sergei Artemov and Elena Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15(6):1059–1073, 2005. 2

[3] Guillaume Aucher. Principles of knowledge, belief and conditional belief. In Manuel Rebuschi, Martine Batt, Gerhard Heinzmann, Franck Lihoreau, Michel Musiol, and Alain Trognon, editors, *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics: Dialogue, Rationality, and Formalism*, pages 97–134, Cham, 2014. Springer International Publishing. 2.5, 2.6

[4] Robert J. Aumann. Agreeing to Disagree. *Annals of Statistics*, 4:1236–1239, 1976. Republished in [?], 2016. 3

[5] A. Baltag, R. Boddy, and S. Smets. *Group Knowledge in Interrogative Epistemology*, pages 131–164. Springer International Publishing, Cham, 2018. 3.2

[6] Alexandru Baltag and Sonja Smets. A Qualitative Theory of Dynamic Interactive Belief Revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games, Vol. 3, pages 9–58. Amsterdam University Press, 2008. 2.6

[7] Johan van Benthem. Epistemic logic and epistemology: The state of their affairs. *Philosophical Studies*, 128(1):49–76, 2006. 1

[8] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, 2001. 2.5, 3, 2.5

[9] Steven Boer and William G. Lycan. *Knowing Who*. Bradford Books / The MIT Press, 1986. 2

[10] Ivan Boh. *Epistemic logic in the later middle ages*. Routledge, 1993. 1

[11] Brian Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980. 2.5

[12] Roderick M Chisholm. The logic of knowing. *The journal of philosophy*, 60(25):773–795, 1963. 4

[13] Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek, and Barteld Kooi, editors. *Handbook of Epistemic Logic*. College Publications, 2015. 1

[14] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, 2007. 1

[15] Ronald Fagin and Joseph Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988. 4

[16] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. The MIT Press, 1995. 1, 2.6, 3, 3.2

[17] Paul Gochet and Pascal Gribomont. Epistemic logic. In D. Gabbay and J. John Woods, editors, *Handbook of the History of Logic, vol. 7*, pages 99–195. Elsevier, 2006. 1

[18] J. Halpern, D. Samet, and E. Segev. Defining knowledge in terms of belief: the modal logic perspective. *The Review of Symbolic Logic*, 3:469–487, 2009. 2.6

[19] Joseph Y Halpern. Should knowledge entail belief? *Journal of Philosophical Logic*, 25(5):483–494, 1996. 2.6

[20] Joseph Y. Halpern and Yoram Moses. Knowledge and Common Knowledge in a Distributed Environment. In *Proceedings of the 3rd ACM Conference on Principles of Distributed Computing*, pages 50–61. ACM, 1984. 3

[21] Vincent F. Hendricks. *Mainstream and Formal Epistemology*. Cambridge University Press, 2006. 2.6

[22] Vincent F. Hendricks and Rasmus K. Rendsvig. Hintikka's *Knowledge and Belief* in Flux. In Hans van Ditmarsch and Gabriel Sandu, editors, *Jaakko Hintikka on Knowledge and Game Theoretical Semantics*, Outstanding Contributions to Logic, pages 317–337. Springer, 2018. 2.6

[23] Vincent F Hendricks and John Symons. Where's the bridge? epistemology and epistemic logic. *Philosophical Studies*, 128(1):137–167, 2006. 1

[24] J. Hintikka. Epistemology without knoweldge and without belief. In *Socratic Epistemology*, pages 11–37. Cambridge University Press, 2007. 1

[25] Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. College Publications, 2nd (2005) edition, 1962. 1, 2, 2.3, 2.6

[26] Jaakko Hintikka. *Semantics for Propositional Attitudes*, pages 21–45. Springer Netherlands, Dordrecht, 1969. 1

[27] Jaakko Hintikka. Impossible possible worlds vindicated. In *Game-Theoretical Semantics*, pages 367–379. Springer, 1979. 4

[28] Jaakko Hintikka and John Symons. Systems of visual identification in neuroscience: Lessons from epistemic logic. *Philosophy of Science*, 70(1):89–104, 2003. 2

[29] Max O. Hocutt et al. Is epistemic logic possible? *Notre Dame Journal of Formal Logic*, 13(4):433–453, 1972. 4

[30] Wesley H. Holliday. Epistemic logic and epistemology. *Handbook of Formal Philosophy*, 2013. 1

[31] Mark Jago. Hintikka and Cresswell on logical omniscience. *Logic and Logical Philosophy*, 15(4):325–354, 2007. 4

[32] Mark Jago. *The Impossible*. Oxford University Press, 2014. 4

[33] Simo Knuuttila. *Modalities in medieval philosophy*. Routledge USA, 1994. 1

[34] Sarit Kraus and Daniel Lehmann. Knowledge, belief and time. In Laurent Kott, editor, *Automata, Languages and Programming*, pages 186–195, Berlin, Heidelberg, 1986. Springer Berlin Heidelberg. 2.6

[35] Daniel Lehmann. Knowledge, Common Knowledge and related puzzles (Extended Summary). *Proceedings of the third annual ACM symposium on Principles of distributed computing (PODC '84)*, pages 62–67, 1984. 2.4, 3

[36] Wolfgang Lenzen. *Recent Work in Epistemic Logic*, volume 30. North Holland Publishing Company, 1978. 2.6

[37] David Lewis. *Convention: A philosophical study*. Harvard University Press, 1969. 3

[38] J.-J.Ch. Meyer and Wiebe van der Hoek. *Epistemic Logic for AI and Computer Science*, volume 41 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1995. 1

[39] John-Jules Meyer. Epistemic logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, pages 183–200. Blackwell Publishers, 2001. 1

[40] Veikko Rantala. Urn Models: A New Kind of Non-Standard Model for First-Order Logic. *Journal of Symbolic Logic*, 4:455–474, 1975. 4

[41] Rasmus K. Rendsvig. Modeling Semantic Competence: A Critical Review of Frege's Puzzle about Identity. In Daniel Lassiter and Marija Slavkovik, editors, *New Directions in Logic, Language and Computation*, pages 140–157. Springer, 2012. 2

[42] Bryan Renne. *Dynamic Epistemic Logic with Justification*. PhD thesis, The City University of New York, 2008. 2

[43] Burkhard C. Schipper. Awareness. In Hans van Ditmarsch, Wiebe Halpern, Joseph Y.and van der Hoek, and Barteld Pieter Kooi, editors, *Handbook of Epistemic Logic*, pages 77–146. College Publications, 2015. 4

[44] Robert Stalnaker. On logics of knowledge and belief. *Philosophical studies*, 128(1):169–199, 2006. 1

[45] Fernando Raymundo Velazquez-Quesada. *Small steps in dynamics of information*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2011. 4

[46] Franz von Kutschera. *Einführung in die intensionale Semantik*. de Gruyter, 1976. 2.6

[47] Georg Henrik von Wright. *An essay in modal logic*, volume 5. North-Holland Publishing Company, 1951. 1

[48] Frans Vorbraak. *As far as I know. Epistemic Logic and Uncertainty*. PhD thesis, Utrecht University, 1993. 2.6

[49] Yanjing Wang. A logic of knowing how. In *International Workshop on Logic, Rationality and Interaction*, pages 392–405. Springer, 2015. 2

[50] Yanjing Wang. Beyond knowing that: a new generation of epistemic logics. In *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 499–533. Springer, 2018. 2

[51] Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, 2000. 2.6