# Epistemic Logic: A Survey

Wiebe van der Hoek [ab] Rineke Verbrugge [c]

[a] Computer Science, Utrecht University, The Netherlands
[b] Computer Science, University of Liverpool, UK
[c] Artificial Intelligence, University of Groningen, The Netherlands

## 1 Introduction

Epistemic logic is the logic of knowledge: how do you reason about the question whether your silent admirer knows that you know that (s)he sent you an anonymous Valentine card? Is it harmful if, at a literature-exam you don't know the contents of a chapter? No, as long as you know that the examiner does not know that you do not know it. Knowing whether your neighbor knows that he regularly plays his radio so loudly that you wake up during the night, may help you to solve the problem in an appropriate way. In negotiations, it will harm you to let the other party know your 'bottom-line', but it may be helpful to disclose other information about yourself, for example about some of your values.

In this article, we will use examples and puzzles to give some flavor of the field and to demonstrate that the notion "it is known that" is meaningful and interesting for researchers in theoretical computer science, artificial intelligence and game theory.

The first person who wrote about epistemic logic was the Swedish-Finnish philosopher G.H. von Wright in his book "An Essay in Modal Logic" [53]. His treatment is completely axiomatic, with no mention of possible semantics. Most philosophical work on epistemic logic following up on Von Wright's work has concentrated on defending certain axioms and denouncing others.

However, the subject of epistemic logic only started to flourish after Kripke's invention of a semantics for modal logic in the early sixties. Kripke introduced a *possible worlds semantics* for modal logics. The name "possible world" is somewhat misleading, because, according to Hintikka [30], "applications to entire universes are scarcely found outside philosophers' speculations. The primary intended applications are to scenarios covering relatively small pieces of space-time". In the context of epistemic logic, one can view worlds that are possible for a certain agent in a world as *epistemic alternatives*, that are compatible with the agent's information in that world. The precise definitions will be given in Section 2. The first full-length book about epistemic logic, Hintikka's "Knowledge and Belief" [29], applies these semantical ideas, although his definitions are not quite the same as the standard ones used today. As Hintikka writes in [30], "the semantics of epistemic logic presents much more interesting problems and solutions than the axiomatic side of the subject". The eighties and early nineties have seen a flurry of activity in the field of epistemic logic. Theoretical computer scientists have, for example, applied it to distributed systems and economists to negotiation. In 1995, this activity culminated when two books about epistemic logic appeared: *Reasoning about Knowledge* by Fagin, Halpern, Moses and Vardi [18] and *Epistemic Logic for AI and Computer Science* by Meyer and Van der Hoek [43].

The most important conference on epistemic logic and related subjects is TARK (since 1996 standing for "Theoretical Aspects of Rationality and Knowledge"), held every other year since 1986 [50]. Newer offshoots that are especially geared to the cooperation of logicians and game theorists are the biannual conference LOFT on "Logic and the Foundations of the Theory of Games and Decisions" [40] and the "International Conference on Logic, Game Theory and Social Choice" [39]. The reader may find further literature about the short but interesting tradition on the interface between logic and game theory in the proceedings of the above-named conferences, as well as in the surveys [7, 12, 47, 51].

Since the middle eighties, there has been much more communication than before between the researchers from different fields using and studying epistemic logic, especially between game theorists and logicians. Game-theorists have sometimes been wary of the heavy logical apparatus introduced by the epistemic logicians: the logicians' results were interesting by themselves, for example showing precisely how agents' knowledge evolves during a game [13]. But how would logic help agents not only to keep track of their own and each other's knowledge and ignorance, but actually to *win* games? Just recently, however, some epistemic logicians are working on precisely the question how to reason about winning strategies for games of imperfect information (see for example [16]).

The rest of this chapter is organized as follows. First, in Section 2, the logics for individual agents in a group are treated. In these, all agents may have different information and thus different epistemic alternatives at each world. Therefore, it is interesting to investigate how they reason about other's knowledge. Example 2.2 is a typical application of epistemic logic in standard computer science. By adding knowledge operators to the language, a program is "derived". This is followed by a short treatment of a logic used for authentication. Then, the semantics and axiomatization of different basic epistemic logics are introduced.

Then, in Section 3, the subject of different types of group knowledge such as common knowledge and distributed knowledge comes to the fore. Even though the notion of common knowledge is intuitively easy, it turns out to be extremely hard to attain or to guarantee this form of group knowledge (see Example 3.1, Example 3.4, Puzzle 3.6, Puzzle 3.8, and Example 3.10). Section 4 describes how agents reason about ignorance. Specifically, the concept of "only knowing" (or "all I know") is treated. All of the previous sections lead to the final one, Section 5 on knowledge and games. Here an example is given of the role of common knowledge in backward induction to find solutions. Then a dynamic-epistemic approach to the evolution of knowledge in games of imperfect information is given, based on the work of Van Ditmarsch. Thus, this chapter is in fact linked to Van Ditmarsch' chapter "The description of game actions in Cluedo" in the present volume.

There is no room here to give further technical and historical background to the examples and puzzles, but the interested reader may find them in the references at the end.

## 2 Knowledge of individual agents in a group

We are now ready to define the formal language for knowledge of individual agents in a group of $m$ agents. The atomic propositions in the language $\mathsf{P}$ typically denote facts about the world, or about a particular game (for instance, an atom can denote that player Alice holds the King of Hearts, cf. Section 5).

**Definition 2.1** Let $\mathsf{P}$ be a non-empty set of propositional variables, and $m \in \mathbb{N}$ be given. The *language* $\mathsf{L}$ is the smallest superset of $\mathsf{P}$ such that

$$\varphi, \psi \in \mathsf{L} \Rightarrow \neg\varphi, (\varphi \wedge \psi), K_i\varphi \in \mathsf{L} \ (i \leq m).$$

We also assume to have the usual definitions for $\vee, \rightarrow$ and $\leftrightarrow$ as logical connectives, as well as the special formula $\perp =_{\text{def}} (p \wedge \neg p)$. In the sequel, we will sometimes use $\Box$ as a variable over the operators $\mathsf{OP} = \{K_1, \ldots, K_m\}$. Indices $i$ and $j$ will range over $\{1, \ldots m\}$.

The intended meaning of $K_i\varphi$ is 'agent $i$ knows $\varphi$'. Thus, for the simplest kind of epistemic logic, where the knowledge of individual agents in a group about the world and the other agents is modeled, it is sufficient to enrich the language of classical propositional logic by unary operators $K_i$. Here, an agent may be a human being, a player, a robot, a machine, or simply a 'process'. Why are these knowledge operators useful? The derivation and correctness proofs of *communication protocols* provide a nice example in computer science.

**Example 2.2 (Alternating bit protocol)** There are two processors, let us say a 'Sender $S$' and a 'Receiver $R$'. The goal is for $S$ to read a tape $X = \langle x_0, x_1, \ldots \rangle$, and to send all the inputs it has read to $R$ over a communication channel. $R$ in turn writes down everything it reads on an output tape $Y$. Unfortunately the channel is not trustworthy, i.e. there is no guarantee that all messages arrive. On the other hand, *some* messages will not get lost, or more precisely: if you repeat sending a certain message long enough, an instance of it will eventually arrive. This property is called *fairness*. Now the question is whether one can write a protocol (or a program) that satisfies the following two constraints, provided that fairness holds:

- *safety:* at any moment, $Y$ is a prefix of $X$;

- *liveness:* every $x_i$ will eventually be written on $Y$.

In the protocol below, $K_S(x_i)$ means that Sender knows that the $i$-th element of $X$ is equal to $x_i$.

PROTOCOL FOR $S$:

$S1$ `i :=0`
$S2$ `while` true do
$S3$     `begin` read $x_i$;
$S4$         send $x_i$ `until` $K_S K_R(x_i)$;
$S5$         send "$K_S K_R(x_i)$" `until` $K_S K_R K_S K_R(x_i)$
$S6$         `i := i + 1`
$S7$     end

PROTOCOL FOR $R$:

$R1$ when $K_R(x_0)$ `set i :=0`
$R2$ `while` true do
$R3$     `begin` write $x_i$;
$R4$         send "$K_R(x_i)$" `until` $K_R K_S K_R(x_i)$;
$R5$         send "$K_R K_S K_R(x_i)$" `until` $K_R(x_{i+1})$
$R6$         `i := i + 1`
$R7$     end

For a simulation of the protocol, the reader be referred to `http://www.ai.rug.nl/mas/protocol/`.

An important aspect of the protocol is that Sender at line $S5$ does not continue reading $X$ and does not yet add 1 to the counter $i$. We will show why this is crucial for guaranteeing safety. For, suppose that the lines $S5$ and $R5$ would be absent, and that instead line $R4$ would read as follows:

$R4'$         send "$K_R(x_i)$" `until` $K_R(x_{i+1})$;

Suppose also, as an example, that $X = \langle a, a, b, \ldots \rangle$. Sender starts by reading $x_0$, an $a$, and sends it to $R$. We know that an instance of that $a$ will arrive at a certain moment, and so by line $R3$ it will be written on $Y$. Receiver then acts as it should and sends an acknowledgement ($R4'$) that will also arrive eventually, thus Sender continues with $S6$ followed by $S3$: once again it reads an $a$ and sends it to Receiver. The latter will eventually receive an instance of that $a$, but will not know how to interpret it: "is this $a$ a repetition of the previous one, because Sender does not know that I know what $x_0$ is, or is this $a$ the next element of the input tape, $x_1$"? This would clearly endanger safety.

As a final remark on the protocol, let us note that it is possible to rewrite the protocol without using any knowledge operators. The result is known as the 'alternating bit protocol'.

Another well-known protocol for sending files is the Transfer Control Protocol (TCP), standardly used on the Internet. A nice feature of the TCP is sliding windows. Instead of dealing with one data item at a time (as the protocol above does), TCP can send a whole window of items at one time. The receiver only needs to acknowledge the last consecutive received package in the window to inform the sender that it has received all the packages in the window thus far. Also, the receiver may determine the size of the window according to its current possibilities. These two aspects reduce the number of acknowledgements and thus the load on the network, while making optimal use of the available bandwidth. A knowledge-based algorithm for TCP is investigated in [48], while a visualization may be viewed at `http://www.ai.rug.nl/~freeks/tcp`.

## 2.1 BAN-logics

Instead of reasoning about whether the contents of the messages arrives safely, another route that one can take is to use epistemic logic to decide whether the receiver of a message can be sure that the sender is really the agent that he purports to be. Also, epistemic logic may be used to guarantee that one knows that intruders don't know specific confidential messages. Here, compared to the analysis of [28] or [48], an additional assumption is that the protocol has to be robust enough to deal with the assumption that the network is hostile. This is for instance mirrored in the technique of *model checking* (cf. [8]) a security protocol, where all possible runs of such a protocol are checked against a (safety or liveness) property. Then, an intruder is often modelled as a process that executes all kinds of attacks at all possible situations in the protocol (cf. [41, 44]).

New and upcoming techniques for the Internet and also wireless telecommunication require or encourage agents to interchange more and more sensitive data, like payment instructions in e-commerce, strategic information between commercial partners, or personal information in, e.g., medical applications. Issues like authentication of the partners in a protocol and the confidentiality of information therefore become of increasing importance: cryptographic protocols are used to distribute keys and authenticate agents and data over hostile networks. Although many of the protocols used look very intricate and hence waterproof, many examples sensitive applications are known to be cracked and then furnished with new, 'improved' protocols.

The application of logical tools to the analysis of security protocols was pioneered by Burrows, Abadi and Needham's [11]. Our brief exposition of this line of research is mainly inspired by and taken from [2]. BAN logic is a modal logic with primitives which describe the beliefs of agents involved in a cryptographic protocol. Using the inference rules of BAN logic, the evolution of the beliefs of agents during a cryptographic protocol can be studied. Here we present some typical rules, explaining the language of BAN in a demand driven way.

The BAN-formalism is built on three sorts of objects: the agents involved in a security protocol, the encryption/decryption and signing/verification keys that the agents possess, and the messages exchanged between agents. The notation $\{M\}_K$ denotes a message encrypted using a key $K$. For a symmetric key $K$ we have $\{\{M\}_K\}_K = M$ for any message $M$, i.e. decrypting with key $K$ a message $M$ that is encrypted with $K$ reveals the contents $M$. For a key pair $\langle EK, DK\rangle$ of a public encryption key $EK$ and a private decryption key $DK$, it holds that $\{\{M\}_{EK}\}_{DK} = M$ for any message $M$. Likewise, for a key pair $\langle SK, VK\rangle$ of a private signing key $SK$ and a public verification key $VK$ it holds that $\{\{H\}_{SK}\}_{VK} = H$ for any hash value $H$. Hash values are obtained by application of a one-way collision-free hash-function. Such a function is supposed to be one-way (given $h(m)$, it is infeasible to compute $m$), and collision-free (no two different messages $m_1$ and $m_2$ have the same hash-value).

In BAN, we have a number of operators describing the beliefs of agents, for which the usual modal properties described in subsection 2.3 apply, like $P$ `believes` $(A \rightarrow B) \rightarrow (P$ `believes` $A \rightarrow P$ `believes` $B)$. On top of that we have the operators `sees` and `possesses` (cf. [11, 22]), denoting that a message is received and that it is in possession of a given agent, respectively. The following rules illustrate some of the authentication and encryption rules:

**(1)** $P$ `believes` $\texttt{secret}(K, P, Q) \wedge P$ `sees` $\{X\}_K$
$\rightarrow P$ `believes` $Q$ `said` $X$

**(2)** $P$ `believes` `belongs_to` $(VK, Q) \wedge P$ `sees` $\{X\}_{SK}$
$\rightarrow P$ `believes` $Q$ `said` $X$

**(3)** $P$ `believes` $\texttt{fresh}(X) \wedge P$ `believes` $Q$ `said` $X$
$\rightarrow P$ `believes` $Q$ `believes` $X$

**(4)** $P$ `possesses` $DK \wedge P$ `sees` $\{X\}_{EK} \rightarrow P$ `sees` $X$

**(5)** $P$ `believes` $\texttt{fresh}(X) \wedge P$ `sees` $(X, Y)$
$\rightarrow P$ `believes` $\texttt{fresh}(X, Y)$

Intuitively, **(1)** says that if an agent $P$ believes that it shares the symmetric key $K$ with an agent $Q$, and agent $P$ receives a message $X$ encrypted under $K$, then agent $P$ believes that agent $Q$ once said message $X$. This rule addresses symmetric encryption. Similarly, **(2)** models digital signatures. If an agent $P$ believes that the verification key $VK$ belongs to an agent $Q$, then $P$ concludes, confronted with a message or hash encrypted with the corresponding signing key $SK$, that the message or hash originates from the agent $Q$. Regarding rule **(3)**, if an agent $P$ believes that certain information is new, i.e. constructed during the current protocol run, and $P$ furthermore believes that $Q$ conveyed this information, then $P$ concludes that the agent $Q$ believes himself this information. (Underlying this is the overall assumption in BAN logic of honesty of agents; the participating agents' behavior is consistent with the particular protocol.) According to **(4)**, if an agent $P$ sees an encrypted message and $P$

possesses the decryption key, then $P$ can read the message itself. Finally, **(5)** is about the freshness of messages. This predicate is used to express that a message has not been seen at the network before, and thus guarantees that the message is not subject to a replay by some third agent. Rule **(5)** then says, that if part of a messages is fresh, the whole message is.

Analysis of an authentication protocol using BAN-logic consists of four phases: (1) first, the initial beliefs of the participating agents are formulated; (2) the protocol security goals are formulated; (3) the effect of the messages of the protocol is formalized in BAN and (4) finally, the final beliefs are shown to fulfill the goals.

As a concluding remark we mention that currently, the semantics of BAN logic is under debate (cf. [10, 1, 49, 54]). At present, we cannot claim that the rules in many BAN-logics are sound or complete. On the one hand, this questions the impact of the derived results (what does it mean that some string has been derived?), but on the other hand, strengthens the call for an adequate model.

## 2.2 Intensional logic

Now let us move to a more formal treatment of the logics of knowledge for individual agents within a group. Specific group phenomena such as common knowledge and distributed knowledge are described in Section 3. First we will describe what distinguishes intensional logics in general from classical logic. Then we move to Kripke semantics for epistemic logics, after which we describe some possible axiomatizations embodying desired properties of epistemic reasoning.

Intensional logic, and in particular epistemic logic, have proven to be popular when modelling informational attitudes of agents. To explain this, let us consider a characteristic property of classical logic, one that intensional logic typically wants to avoid.

**Observation 2.3 (Extensionality)** Let $[q/p]\varphi$ denote the formula $\varphi$, but with (an arbitrary number of occurrences of) the subformula $p$ replaced by $q$. Then, classical logic encompasses the following property:

$$\models (p \leftrightarrow q) \rightarrow (\varphi \leftrightarrow [q/p]\varphi)$$

In words, extensionality says that, to determine the truth-value of a formula $\varphi$, it is only the truth-value of its subformulas that counts: if we replace any occurrence of a subformula $p$ by another formula $q$ *with the same truth-value*, then this does not matter for the value of the formula as a whole. Since the truth-value of a formula is sometimes also denoted as its extension, we can rephrase Observation 2.3 loosely as: the extension of a complex formula is determined by the extension of its subformulas, not by their form.

To give an example, let $p$ denote that the Society of Dynamic Games has its office in Saint-Petersburg, and let $q$ be the statement that it has a homepage on the web. We then quickly recognize that $p$ and $q$ are both true, and thus are equivalent in the present situation, even if they are not logically equivalent: we have $(p \leftrightarrow q)$. Furthermore, let $l$ denote that logic is important (true) and $w$ that the Society is situated in Moscow (false). Then, according to extensionality, we have that $(w \rightarrow q)$ is equivalent to $(w \rightarrow p)$, and $l \vee (q \wedge w)$ is equivalent to $l \vee (p \wedge w)$. Combining complex assertions and then calculating their truth-value, is done by substituting the values (extension) for the subformulas. $\mathrm{Ext}(l \vee (q \wedge w)) == \mathrm{Ext}(l \vee (p \wedge w))$

Having established that classical logic satisfies the property of extensionality, one may wonder whether this is desirable, or whether there are constructs in natural language that do not satisfy this principle. It appears there are many. Let $c$ denote that the Society holds office in Russia. Then '$c$, because $p$' is obviously true, whereas '$c$, because of $q$' makes no sense. Noting that, by extensionality, $p \rightarrow c$ and $q \rightarrow c$ are equivalent, we obtain two conclusions: '$B$ because of $A$' cannot be modelled by $A \rightarrow B$, and, even stronger, 'because of' cannot be modelled in propositional logic at all (since 'because of' is not extensional, whereas classical logic is). This observation was in fact one of the motivations to develop modal logic (see [37, 17]).

But there are more examples of constructs that are not extensional. For instance 'I wish the Society had a home-page' is not the same as 'I wish the Society had its office in Saint-Petersburg'. Also, knowing $p$ is different from knowing $q$. Compare 'five years ago, $p$', with 'five years ago, $q$'. 'When moving office out of town, $\neg p$' does not necessarily mean 'when moving office out of town, $\neg q$'. Thus, when reasoning with motivational attitudes (wishing), informational attitudes (knowing), temporal properties (five years ago) or hypothetical events (when moving), we don't have extensionality. Ergo: we cannot use classical logic to deal with them.

### 2.3 Semantics

Epistemic logic is one attempt to circumvent these problems and to accurately model situations in which extensionality fails for informational attitudes.

The intuition of formulas in the epistemic language $\mathsf{L}$ (see definition 2.1) is best explained by looking at the semantics of $K_i\varphi$. Given a situation $s$ (for the moment, think of it as a truth-assignment to atoms), $K_i\varphi$ is true in $s$ if $\varphi$ is true in all situations $t$ that agent $i$ views as *epistemic alternatives* for $s$, compatible with the information of agent $i$ at world $s$. For instance, if $p$ denotes 'it is sunny in Saint-Petersburg' and $q$ 'it is sunny in Amsterdam', and $s$ is the situation where I am in Saint-Petersburg, where it is sunny, then for me, two situations are relevant, if compatibility with my information is concerned: $t_1$ in which $p \wedge q$ is true, and $t_2$ in which $(p \wedge \neg q)$ is. Since in all my alternatives $p$ is true, I know $p$ (denoted by $K_ip$), but we also have $\neg K_iq$ and $\neg K_i\neg q$ (see Figure 1). Note that this perfectly solves our problem of extensionality: in $s$, we have $(p \leftrightarrow q)$ but also $(K_ip \wedge \neg K_iq)$.

We are now ready to define the semantics for our modal language $\mathsf{L}$ formally: a Kripke model $M$ is a tuple $M = \langle S, \pi, R_1, \ldots, R_m \rangle$ where $S$ is a non-empty set of worlds or states or situations $s$, $\pi$ computes, for every state $s$ the truth-value

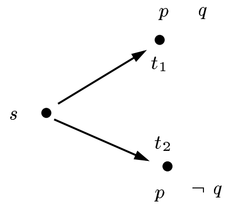


Figure 1: A Kripke model for knowledge without extensionality

$\pi(s)(p)$ for every atom $p$, and each $R_i$ for $i \leq m$ is a binary accessibility relation between worlds. In order to determine whether a formula $\varphi \in \mathsf{L}$ is true in $w$, denoted by $(M, w) \models \varphi$, we look at the structure of $\varphi$:

| | | |
|---|---|---|
| $M, s \models p$ | iff | $\pi(s)(p) = true$ |
| $M, s \models (\varphi_1 \wedge \varphi_2)$ | iff | $M, s \models \varphi_1$ and $M, s \models \varphi_2$ |
| $M, s \models \neg\varphi$ | iff | not $M, s \models \varphi$ |
| $M, s \models K_i\varphi$ | iff | for all $t$ such that $R_ist, M, t \models \varphi$ |

Under such a definition, we say that $K_i$ is the necessity operator for an accessibility relation $R_i$. The clause for $K_i\varphi$ is sometimes also written in a functional way: $\forall t \in R_i(s), M, t \models \varphi$. A formula $\varphi$ is true in a model, written $M \models \varphi$, if $M, s \models \varphi$ for all $s \in S$. If $\mathcal{M}$ is a class of models, $\varphi$ is said to be *valid on* $\mathcal{M}$, if for all $M \in \mathcal{M}, M \models \varphi$. Restricting validity to classes of models is a valuable notion in modal logic, as we shall see shortly.

It is an interesting question to determine what the properties of any modal logic are, i.e., properties $\varphi$ that are valid in every Kripke model. For such $\varphi$, we write $\models \varphi$.

**Definition 2.4** Let $\varphi, \psi$ be formulae in $\mathsf{L}$, and let $K_i$ be an epistemic operator for $i \leq m$. Then the following hold:

- $\models K_i\varphi \wedge K_i(\varphi \rightarrow \psi) \rightarrow K_i\psi$ $\quad LO1$
- $\models \varphi \Rightarrow \models K_i\varphi$ $\quad LO2$
- $\models \varphi \rightarrow \psi \Rightarrow \models K_i\varphi \rightarrow K_i\psi$ $\quad LO3$
- $\models \varphi \leftrightarrow \psi \Rightarrow \models K_i\varphi \leftrightarrow K_i\psi$ $\quad LO4$
- $\models (K_i\varphi \wedge K_i\psi) \rightarrow K_i(\varphi \wedge \psi)$ $\quad LO5$
- $\models K_i\varphi \rightarrow K_i(\varphi \vee \psi)$ $\quad LO6$
- $\models \neg(K_i\varphi \wedge K_i\neg\varphi)$ $\quad LO7$

For example, $LO1$ says that knowledge is closed under consequences. $LO2$ expresses that agents know all validities. The properties of Definition 2.4 reflect *idealized* notions of knowledge, that do not necessarily hold for human beings.

For example, many people do not know all tautologies of propositional logic, so $LO2$ does not hold for them.

We will see later how, in many systems, these properties are nevertheless accepted; for a discussion on weakening these properties we refer to [18, 43]. The fact that the above properties hold in all Kripke models is sometimes also referred to as the problem of *logical omniscience*, since they express that agents are omniscient, perfect logical reasoners.

The definition of Kripke semantics can easily be adapted to other modal logics, by replacing the $K_i$ by other modal operators. In a nutshell, the modal language adds one or more unary operators $\Box$ to the language, where $\Box\varphi$ can be used to model '$\varphi$ is believed', or '$\varphi$ is always the case', '$\varphi$ is a desire', '$\varphi$ is obligatory', '$\varphi$ will always hold from now on', or '$\varphi$ is a result of executing program $\pi$'. The interpretation of the accessibility relations is adapted appropriately for each instance.

## 2.4 Axiomatization

The following definition establishes the exact properties of the notion $K_i$ for $i \leq m$. This is the minimal axiomatic system $\mathcal{K}_m$, corresponding to the truths that hold in all Kripke models.

**Definition 2.5** *The basic epistemic logic $\mathcal{K}_m$, where we have an operator $K_i$ for every $i \leq m$, is comprised of the axioms $A1, A2$ below, and the derivation rules $R_1$ and $R_2$. The corresponding axioms are the following:*

$A1$ *any axiomatization for propositional logic*

$A2$ $(K_i\varphi \wedge K_i(\varphi \to \psi)) \to K_i\psi$

*On top of that, we assume the following derivation rules:*

$R1$ $\vdash \varphi, \vdash \varphi \to \psi \ \Rightarrow \vdash \psi$

$R2$ $\vdash \varphi \ \Rightarrow \vdash K_i\varphi$*, for all* $i \leq m$

Other well-known axiom systems, stronger than the minimal system, have been defined. In each of them, in addition to $\mathcal{K}_m$, a choice is made from among the following axioms:

$A3$ $K_i\varphi \to \varphi$

$A4$ $K_i\varphi \to K_iK_i\varphi$

$A5$ $\neg K_i\varphi \to K_i\neg K_i\varphi$

$D$ $\neg(K_i\varphi \wedge K_i\neg\varphi)$

For historical reasons, some of the best-known systems have been given the following names:

$\mathcal{T}_m = \mathcal{K}_m + A3$

$\mathcal{S}4_m = \mathcal{T}_m + A4$

$\mathcal{S}5_m = \mathcal{S}4_m + A5$

$\mathcal{KD}45_m = \mathcal{K}_m + D + A4 + A5$

For any of these systems $\mathcal{X}$, a formula $\varphi$ is *derivable* from $\mathcal{X}$, notation $\vdash_{\mathcal{X}} \varphi$, if there is a proof for $\varphi$ that only uses the axioms and rules of $\mathcal{X}$.

Thus, the axioms and rules of $\mathcal{K}_m$ are assumed to hold for all rational agents, where 'rational' means that they are logically omniscient; it is not assumed that agents maximize their utilities. That the agents are taken to be rational, is perhaps best reflected by the fact that we have the property $K_i\varphi \wedge K_i(\varphi \rightarrow \psi)) \rightarrow K_i\psi$ and rule R1, reflecting the semantic properties $LO1$ and $LO2$ discussed in Definition 2.4. Note that these properties are part and parcel of the modal approach to epistemic logic.

Furthermore, knowledge is assumed to be *veridical*, in the sense that agents do not know falsities (see $A3$). Agents are also assumed to have *positive introspection* (if they know something, they know that they know it, see A4) and *negative introspection* (if they do not know something, they know that they do not know it, see $A5$). Even though $A4$ and, in particular, $A5$ are somewhat controversial for human knowledge, the full system $\mathcal{S}5_m$ is often considered as *the* standard epistemic logic.

Note that, if $K_i\varphi$ were to be interpreted as 'agent $i$ believes $\varphi$' it is not reasonable to assume $A3$. Instead, the weaker axiom $D$ is often assumed to hold: agents do not believe any inconsistensies. A well-known logic for beliefs (or *doxastic logic*) is $\mathcal{KD}45_m$.

One of the attractive features of modal logic is that the each of the axioms that we mentioned above is immediately reflected in a structural property on the semantics. To wit, if one chooses for a modal logic satisfying $A3$, one has to stipulate *reflexivity* ($\forall s R_i ss$) of the accessibility relation $R_i$. In the same spirit, $A4$ corresponds to transitivity ($\forall s, t, u((R_i st \& R_i tu) \rightarrow R_i su)$), $A5$ to Euclidicity ($\forall s, t, u((R_i st \& R_i su) \rightarrow R_i tu)$) and axiom $D$ to seriality ($\forall s \exists t R_i st$). In this way, one obtains a modular tool to build modal systems: the system $\mathcal{S}5_m$ for example, is syntactically obtained by adding the axioms $A3, A4$ and $A5$ to $\mathcal{K}_m$, and the appropriate models are obtained by combining the corresponding constraints: reflexivity, transitivity and Euclidicity. Since the combination of the latter three properties yields an *equivalence relation*, we can phrase the relevant property that couples semantics to axiomatization, for $\mathcal{S}5_m$, in the following way. Let $\mathcal{EQ}$ be the class of models $M = \langle S, \pi, R_1, \ldots, R_m \rangle$ for which each $R_i$ is an equivalence relation. Then, the theorems *derivable* in the axiom system $\mathcal{S}5$ are exactly the validities of the class $\mathcal{EQ}$:

$$\text{for all } \varphi : \mathcal{EQ} \models \varphi \iff \vdash_{\mathcal{S}5_m} \varphi$$

## 2.5 Relation with Aumann's definitions

Readers familiar with game theory may be best acquainted with epistemic notions as defined by Aumann, as summarized in his [4]. In his terminology, our set of states $S$ is a space $\Omega$ of states of the world. Second, Aumann assumes *partitions* $\mathcal{F}_i$ on this set $\Omega$, one for each agent. As a reminder, a partition of a set $S$ is a set $P$ whose elements are non-empty disjoint subsets of $S$ whose union is equal to $S$. Given a state $\omega$, and a partition $\mathcal{F}_i$, all the states that are in the same block of the partition (called 'atom' by Aumann) as $\omega$ are the states that the agent $i$ cannot distinguish from $\omega$, and, hence, from each other. This approach is equivalent to our definition, when one accepts $\mathcal{S}5_m$ as the system for knowledge. For, since the accessibility relations $R_i$ in models for $\mathcal{S}5_m$ are equivalence relations, they uniquely determine a partition $\mathcal{F}_i$, for each agent $i$.

Aumann also defines *events*, rather than *formulas*. An event $E$ is just a subset of the set of states $\Omega$ (one can think of this set as comprising those states in which '$E$ is true'). Then, union of events corresponds to disjunction of propositions, intersection to conjunction, complement to negation and the subset-relation to implication. Thus, when $E \subseteq F$, Aumann says that '$E$ entails $F$'. Mathematically, the set of states that $i$ considers possible in $\omega$ is denoted by $\mathbf{I}_i(\omega)$. The property of veridicality (axiom $A3$) for instance, is then guaranteed by the constraint $\omega \in \mathbf{I}_i(\omega)$, for instance. The set $\{\mathbf{I}_i(\omega) | \omega \in \Omega\}$ then is required to be a partition: it is the *information partition* of agent $i$, and each member $\mathbf{I}_i(\omega)$ is called an *information set*. The family of all unions of events in $\mathcal{F}_i$ is denoted $\mathcal{K}_i$ and called the *universal field*; here, all the events receive their interpretation, and hence it is assumed to be closed under arbitrary union and complementation. Knowledge is now defined as an event as well: if $E$ is an event then $K_iE$ is a new event, defined as follows:

1. $\omega \in K_iE$ iff $\mathbf{I}_i(\omega) \subseteq E$
2. $K_iE$ is the largest element of $\mathcal{K}_i$ that is included in $E$.

In the words of Aumann:

> ... $K_iE$ is the event that $i$ knows that the event $E$ obtains (...); more explicitly, the set of all states $\omega$ at which $i$ knows that $E$ contains $\omega$ (he usually will not know the true $\omega$)

For further details on this 'alternative' approach to epistemic systems, we refer to Aumann's [4]. There, one also finds a discussion about the question whether, in what sense and why the space $\Omega$ and the partitions $\mathcal{F}_i$ can be taken as given and commonly known by the players involved in a game.

# 3 Group epistemics

In this section, we introduce some notions of group knowledge for multiple agents that are relevant for both game theory and computer science: 'everyone knows', common knowledge and implicit knowledge within a group.

To start with, for a group of $n$ agents $\{1, \ldots, n\}$, one can define 'Everyone Knows' ($E\varphi$) by $E\varphi \equiv K_1\varphi \wedge \cdots \wedge K_n\varphi$. A natural question now would be: Given that all agents are positively introspective, is then the notion of $E$-knowledge as well? We would expect not: consider 20.000 fans of the Rolling Stones, all having positive introspective knowledge, and all going to the same concert. Each fan knows that the concert starts at ten ($t$) so we have $K_it$, $K_iK_it$ and $Et$. But there is no reason to assume that $EEt$ holds: how should John know that Mary (whom he does not know) knows that the concert starts at ten?

A more intriguing notion of group knowledge is 'Common Knowledge' ($C\varphi$), that should mean something like $E\varphi \wedge EE\varphi \wedge EEE\varphi \wedge \ldots$. (Unfortunately, such an infinite conjunction is not allowed in the language of epistemic logic.) This notion is rather crucial to game theorists. We quote Aumann commenting on the assumption in game theory that rationality of the players, the rules of the game, and the set of players are commonly known ([3, p. 31]):

> The common knowledge assumption underlies all of game theory and much of economic theory. Whatever be the model under discussion, whether complete or incomplete information, consistent or inconsistent, repeated or one-shot, cooperative or non-cooperative, the model itself must be assumed common knowledge; otherwise the model is insufficiently specified, and the analysis incoherent.

One can intuitively grasp the fact that the number of iterations of the $E$-operator makes a real difference in practice. For example, suppose that $\varphi$ stands for "Saint Nicholas does not exist"[1].

Imagine the difference in how a Dutch family's celebration of Saint Nicholas' Eve would look like if $K_1\varphi \wedge \neg E\varphi$ holds, as compared with the situation where $E\varphi \wedge \neg EE\varphi$ holds, or the one where $EE\varphi \wedge \neg EEE\varphi$ holds.

The notion of *common knowledge* arises from David Lewis' *Convention: A Philosophical Study* [38]. One of the questions in his book is about the convention of driving on a certain side of the road. What kind of knowledge is needed for every driver to feel reasonably safe? Suppose that all Russian drivers drive on the right side of the road. That fact by itself is not enough to make all drivers feel safe: it seems necessary that "everybody knows that everybody drives on the right side". Now imagine the strange situation where everyone drives on the right because they know that all others drive on the right, but that everyone holds the following false belief: "except for myself, everyone else drives on the right just by habit, and would continue to do so no matter what he expected others to do". Lewis argues that in this imaginary situation one cannot really say that there is a convention to drive on the right. After giving some more complex imaginary examples, Lewis proposes that if there is a convention among a group that $\varphi$, then everyone knows $\varphi$, everyone knows that everyone knows $\varphi$, everyone knows that everyone knows that everyone knows $\varphi$, and so on ad infinitum. In such a case, we say that the group has common knowledge of $\varphi$.

To show the importance of common knowledge in everyday conversations, consider the following situation. Suppose a friend asks you "Did you go to the concert?", referring to the Rolling Stones' concert in the City Stadium last June. Of course to understand each other you and your friend must both know that "the concert" refers to the Rolling Stones' concert in the City Stadium last June, but also you must know that both of you know it (so that they will know that your answer is appropriate to your friend's question), you must know that

[1] On December 5, according to tradition, Saint Nicholas is supposed to visit Dutch homes and to bring presents. Children generally start to disbelieve in his existence when they are around six years old, but for various reasons many children like to pretend to believe in him a little while longer.

both of you know that both of you know it (so that you will know that your friend's response to your answer is appropriate), and so on.

Another way to grasp the notion of common knowledge is to realize in which situations $C\varphi$ does *not* hold for a group. This is the case as long as someone, on the grounds of their knowledge, holds it for a possibility that someone holds it for a possibility that someone ... that $\varphi$ does not hold. The following example illustrates such a situation.

**Example 3.1 (Alco at the conference)** Alco is visiting a conference in Barcelona, where at a certain point during the afternoon he becomes bored and decides to lounge in the hotel bar. While he is enjoying himself there, an important practical announcement $\varphi$ is made in the lecture room. Of course at that moment $C\varphi$ does not hold, nor even $E\varphi$. But now suppose that in the bar the announcement comes through by way of an intercom connected to the lecture room. Then we do have $E\varphi$, but not $C\varphi$; after all, the other visitors of the conference do not know that Alco knows $\varphi$.

After hearing $\varphi$, Alco leaves the hotel for some sightseeing in the city. At that moment someone in the lecture room worriedly asks whether Alco knows $\varphi$, upon which the program chair reassures her that this is indeed the case, because of the intercom. Of course at that moment, $C\varphi$ still does not hold!

The example above illustrates that common knowledge is a very strong notion, which therefore holds only for very weak propositions $\varphi$.

The notion of *implicit* or *distributed knowledge* also helps to understand processes within a group of people or collaborating agents. Distributed knowledge is the knowledge that is implicitly present in a group, and which might become explicit if the agents were able to communicate. For instance, it is possible that no agent knows the assertion $\psi$, while at the same time $D\psi$ may be derived from $K_1\phi \wedge K_2(\phi \rightarrow \psi)$. Suppose that you know that all students of modal logic are at least 19 years old, and I know that Kripke is 17 years old, then together we have distributed knowledge that Kripke is not a student of modal logic. In general, we have distributed knowledge of $\varphi$ if, by putting our knowledge together, $\varphi$ may be deduced, even if none of us individually knows $\varphi$. (Actually, Kripke was 17 when he invented what is now called Kripke semantics.) A common example of distributed knowledge in a group is, for instance, the fact whether two members of that group have the same birthday.

## 3.1 Language and semantics

Now we are ready to give the definition of the full language of epistemic logic, including the group notions.

**Definition 3.2** Let $\mathsf{P}$ be a non-empty set of propositional variables, and $m \in \mathbb{N}$ be given. The *language* $\mathsf{L}'$ is the smallest superset of $\mathsf{P}$ such that

$$\varphi, \psi \in \mathsf{L}' \Rightarrow \neg\varphi, (\varphi \wedge \psi), K_i\varphi, E\varphi, C\varphi, D\varphi \in \mathsf{L}' \ (i \leq m).$$

Thus, $\mathsf{L}'$ extends $\mathsf{L}$ of definitiom 2.1 with the operators $E, C$, and $D$.

Here, $E\varphi$ has to be read as 'everyone knows $\varphi$' and $C\varphi$ is 'it is common knowledge that $\varphi$'. Moreover, $D\varphi$ means '$\varphi$ is distributed knowledge', or '$\varphi$ is implicit knowledge of the $m$ agents'[2].

The semantics for our modal language $\mathsf{L}'$ is very similar to the case for individual knowledge: a Kripke model $M$ is again a tuple $M = \langle S, \pi, R_1, \ldots, R_m \rangle$ as defined there. We only need to add clauses for the new operators. For $E$, this is simple:

$$M, s \models E\varphi \text{ iff for all } i \leq m,\ M, s \models K_i\varphi,$$

For defining the meaning of the $C$-operator, additional work needs to be done. Define $t$ to be *reachable* from $s$ if there is a path in the Kripke model from $s$ to $t$ using pairs from possibly different $R_i$'s that are associated with agents $i \leq m$. Then the following property holds:

$$M, s \models C\varphi \text{ iff } M, t \models \varphi \text{ for all } t \text{ that are reachable from } s.$$

Finally, the semantics of $D$ is defined as follows:

$$M, s \models D\varphi \text{ iff } M, t \models \varphi \text{ for all } t \text{ such that } (s, t) \in R_1 \cap \ldots \cap R_m.$$

In [33], the authors ask themselves whether, given a $\mathcal{S}5_m$ Kripke model, distributed knowledge can always be made explicit by communication. The answer is strikingly simple and natural: only if the underlying Kripke model is *finite* and *distinguishing*, which means that for all states $s$ and $t$ there is a formula $\varphi_s$ such that $M, s \models \varphi_s$, but $M, t \not\models \varphi_s$. Their point is easily illustrated.

**Example 3.3** Consider the model of Figure 2, which is an $S5_2$-model, in which reflexive arrows are not drawn. Let us suppose that the dashed arrows denote the accessibility relation of agent 1, and the solid arrows denote those of agent 2. It is easily verified that the states $x$ and $x'$ verify the same formulas, as do $z$ and $z'$. For instance, we have $M, x \models \neg K_1 q \wedge \neg K_2 q$ and $M, x' \models \neg K_1 q \wedge \neg K_2 q$. Even stronger, one proves by induction on formulas $\varphi$ that $M, x \models K_1\varphi \Leftrightarrow M, x \models K_2\varphi$. But we also have $M, x \models Dq$, since $(x, x)$ is the only pair $(x, \cdot)$ in $R_1 \cap R_2$. It is clear that $q$ will never become explicit knowledge by one of the agents if they communicate, since they already know the same facts beforehand.

Concerning the relations between the type of knowledge presented, one may observe that we have:

$$C\varphi \Rightarrow E\varphi \Rightarrow K_i\varphi \Rightarrow D\varphi \Rightarrow \varphi$$

[2]The term "implicit knowledge" may be confusing because of the distiction between explicit and implicit knowledge appearing in awareness logics [19]. From now on we only talk about distributed knowledge.

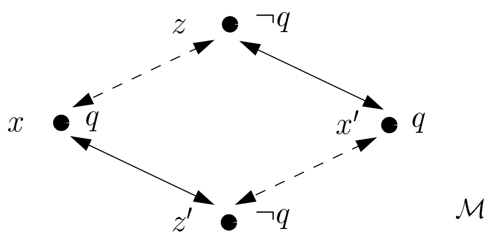


Figure 2: A finite model for two agents

So, common knowledge is the strongest notion, and distributed knowledge the weakest. But this also means that common knowledge will, in general, only be obtained about very weak statements $\varphi$, where distributed knowledge may be obtained from statements $\varphi$ that are not known by anybody. This is why Halpern and Moses in their famous [26] rephrase $C\varphi$ as 'any fool knows $\varphi$' and $D\varphi$ as 'a wise man knows $\varphi$'.

Turning back to common knowledge, we are ready to show a well-known example.

**Example 3.4 (The muddy children)** In this example [3] the principal players are a father and $k$ children, of whom $m$ (with $m \leq k$) have mud on their forehead. The father wants to have a serious talk with the muddy children. Thus, he calls all the children together. None of them knows whether it is muddy or not, but they can all accurately perceive the other children and judge whether they are muddy. Moreover, all this is general knowledge; it is also common knowledge that all children are perfect logical reasoners and have even successfully finished a course on epistemic logic. Now father has a very simple announcement $\psi$ to make:

> At least one of you is muddy. If you know that you are muddy, please come forward.

After this, nothing happens (except in case $m = 1$). When the father notices this, he literally repeats the announcement $\psi$. Once again, nothing happens (except in case $m = 2$). The announcement and subsequent silence are repeated until the father's $m$-th announcement. Suddenly all $m$ muddy children step forward! It would go too far to explain the logical techniques needed to give a sound explanation, but one gets a good idea when investigating what happens in the cases $m = 1, 2$. Thus, suppose $m = 1$ and father just announced $\psi$, then the only muddy child knows it is muddy, because it does not see any muddy companions. It obediently steps forward. Now suppose $m = 2$, and call the muddy children $m_1$ and $m_2$. Let us follow $m_2$'s reasoning. After the first '$\psi$', $m_2$ reasons about $m_1$ just like we did in the previous case: "I don't know whether I'm muddy. If not, $m_1$ wouldn't see any muddy companions

[3] This one is a more politically correct version of the folklore 'cheating husbands problem'.

and would step forward". At the father's second '$\psi$', $m_2$ knows that $m_1$ has not in fact stepped forward, so:"$m_1$ must have seen another muddy child. I don't, so that must have been me". Now $m_2$ steps forward, and $m_1$ as well (by a symmetrical argument). Note finally, that even if many children are muddy, there is no common knowledge that there is even at least one muddy child before the father makes his first announcement! For example, in case $m = 2$, child $m_1$ holds it to be possible that it is not muddy and that simultaneously $m_2$ holds it as possible that $m_2$ is not muddy either.

Let us analyze the muddy children problem semantically, where we have three children, creatively named 1, 2 and 3. In Figure 3, the situation just before the father's first announcement is modeled; worlds are denoted as triples $(x, y, z)$. The world $(1, 1, 0)$ for instance denotes that child 1 and 2 are muddy, and 3 is not. Given the fact that every child sees the others but not itself, we can understand that child 1 'owns the solid lines' in the figure, since 1 can never distinguish between two states $(0, y, z)$ and $(1, y, z)$. Similar arguments apply to children 2 and 3.

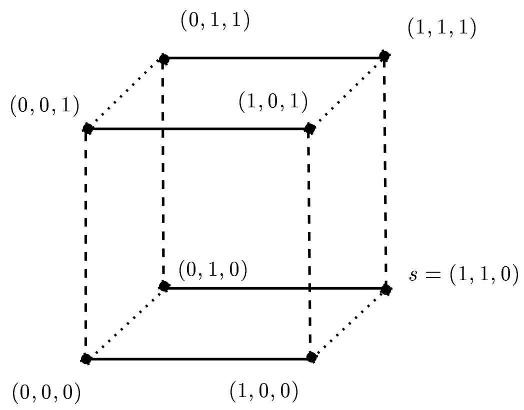


Figure 3: Kripke structure $M$ representing the puzzle in the case where 2 out of 3 children are muddy, just before the father's announcement

Let us consider what epistemic truths we have in the state $(M, s)$, with $s = (1, 1, 0)$. The only propositional atoms we use are $m_i$ $(i = 1, 2, 3)$ with meaning 'child $i$ is muddy'. In state $s$, we then have

($a$) $M, s \models \neg(K_1 m_1 \vee K_1 \neg m_1)$ (child 1 does not know whether it is muddy);

($b$) $M, s \models K_1 m_2 \wedge K_1 \neg m_3$ (child 1 knows that 2 is muddy, but that 3 is not);

($c$) $M, s \models K_1(m_2 \wedge \neg K_2 m_2) \wedge K_1(\neg m_3 \wedge \neg K_3 \neg m_3)$
child 1 knows that 2 is muddy without knowing it, and also that 3 is mudless without knowing that;

($d$) $M, s \models K_1(m_1 \rightarrow (K_2 m_1 \wedge K_2 K_3 m_1))$

child 1 knows that, if he is muddy, 2 knows it, and that 2 then also knows that 3 knows it.

Regarding group notions, we observe the following, in $s$. Let $\psi(j)$ denote that at least $j$ children are muddy (which can be represented as a disjunction of conjunctions of $(\neg)m_i$'s):

$(e)$ $M, s \models E\psi(1) \wedge \neg Em_1 \wedge \neg Em_2 \wedge \neg Em_3$
everybody knows that there is at least one muddy child, but nobody is known by everybody to be muddy

$(f)$ $M, s \models K_3E\psi(1) \wedge \neg K_2E\psi(1)$
child 3 knows that everybody knows that there is at least one muddy child, but child 2 does not know that everybody knows at least one child to be muddy. To see the second conjunct, note that $M, (1,0,0) \models \neg E\psi(1)$, hence $M, s \models \neg K_2E\psi(1))$.

$(g)$ $M, s \models \neg C\psi(1)$
it is not common knowledge that there is at least one muddy child! This follows immediately from the previous item, but also directly from the model: one can find a path from $s = (1,1,0)$ via $(0,1,0)$ to $(0,0,0)$, the latter state being one in no child is muddy.

We leave an analysis of the situations that may arise after the father's subsequent announcements to the reader. At any rate, after the father announces $\psi$, the Kripke structure of Figure 3 is truncated: the world $(0,0,0)$ becomes inaccessible for all three children.

## 3.2 Axiomatization

The following definition establishes the exact properties of and relations between the notions $E, C$ and $D$ introduced in the previous section.

**Definition 3.5** *We define a number of epistemic logics. The basic one being $S5_m$, where we only have an operator $K_i$ for every $i \leq m$. The logic $S5_m(CDE)$ (or L for short) adds to $S5_m$ all the axioms concerning the operators $C, D$ and $E$ below. Intermediate systems are understood from their notation: the logic $S5_m(DE)$ for instance adds axioms A6 and A11 − A15 to $S5_m$. The relevant extra axioms are the following:*

A6 $E\varphi \leftrightarrow (K_1\varphi \wedge \cdots \wedge K_m\varphi)$

A7 $C\varphi \rightarrow \varphi$
A8 $C\varphi \rightarrow EC\varphi$
A9 $(C\varphi \wedge C(\varphi \rightarrow \psi)) \rightarrow C\psi$
A10 $C(\varphi \rightarrow E\varphi) \rightarrow (\varphi \rightarrow C\varphi)$

$A11 \quad K_i\varphi \to D\varphi$
$A12 \quad (D\varphi \wedge D(\varphi \to \psi)) \to D\psi$
$A13 \quad D\varphi \to \varphi$
$A14 \quad D\varphi \to DD\varphi$
$A15 \quad \neg D\varphi \to D\neg D\varphi$

*Additionally, we assume the following derivation rule in addition to those of* $S5_m$*:*

$R3 \quad \vdash \varphi \;\Rightarrow \vdash C\varphi$

Axiom $A6$ can be understood as the definition of $E$, whereas $A8$ says that all common knowledge is known by everybody as such. Axiom $A10$ is also known as the *induction axiom*. The axiom explains how one can derive that $\varphi$ is common knowledge: by deriving $\varphi$ itself together with some common knowledge about $\varphi \to E\varphi$.

Notice that the rationality properties of individual agents, as described in Subsection 2.3, carry over to some of the group notions. For example, we have $(\Box\varphi \wedge \Box(\varphi \to \psi)) \to \Box\psi$ for all $\Box \in \{E, C, D\}$ —which follows from $A9, A12$ and, in the case of $E$, with a simple calculation using $A6$ and $A2$.

Common knowledge and distributed knowledge are all supposed to be *veridical* ($A7$ and $A13$, respectively). The properties of *positive* as well as *negative introspection* are also ascribed to distributed knowledge ($A14$ and $A15$, respectively). Both properties of introspection can be shown to hold for common knowledge as well. For negative introspection, one first shows that $S5_m(CDE) \vdash C\varphi \leftrightarrow K_iC\varphi$; we leave the proof to the reader as a nice exercise.

Finally, the rule $R3$ expresses another rationality principle of the (group of) agents we consider: it guarantees that $L$-derivable formulas give rise to the derivability in $L$ of the same formula, prefixed by any of the operators from $\{E, C, D\}$; that is, one easily proves that, for every $\Box \in \{E, C, D\}$: $\vdash \alpha \;\Rightarrow \vdash \Box\alpha$.

For negotiations and games, it is not only important for participants to know what the others do know, but also what the others do *not* know. Thus, your ignorance can provide useful information to other agents (see also Section 4). A well-known example of this phenomenon is the *wise men* puzzle, in which one wise person can derive the color of his hat from the fact that his colleagues have said they do not know the color of their hats. A somewhat more complex variant of this puzzle is the muddy children in example 3.4. The following puzzle, adapted from a paper by John McCarthy [42] but first presented in [20], is less well known.

**Puzzle 3.6 (Sum and Product)** Two persons, $S$ and $P$, find themselves in a room, of which they do not know the dimensions breadth $b$ and length $l$, both integers. $S$ is told (in secret) the sum of the two integers, and $P$ is told (again in secret) their product. It is common knowledge among them that $S$ knows the sum and $P$ the product, as well as the constraint that $2 \leq b \leq l \leq 99$. Moreover it is generally known that both are good at arithmetic and epistemic logic. At this point, the following dialogue arises:

$P$: "I don't know the numbers."
$S$: "I knew you didn't know. I don't know either."
$P$: "Now I know the numbers."
$S$: "Now I know them too."

In view of the above dialogue it is possible to reconstruct $b$ and $l$. What are the numbers? (For a discussion of this problem by two groups of AI students from Groningen, see `http://www.ai.rug.nl/mas/samenpro/`, where a program is given, and `http://www.ai.rug.nl/mas/samenprosem/`, where the problem is solved semantically.)

Without giving away the solution, we will make some remarks about the problem from a semantic epistemic viewpoint. It is clear that the puzzle can be solved by starting with all Kripke models represented by points $(b, l)$ with $2 \leq b \leq l \leq 99$. The two accessibility relations are quite obvious as well:

$$(b_1, l_1)R_S(b_2, l_2) \text{ iff } b_1 + l_1 = b_2 + l_2$$

$$(b_1, l_1)R_P(b_2, l_2) \text{ iff } b_1 \cdot l_1 = b_2 \cdot l_2$$

Now, according to the first statement by $P$, we may remove all points that are only $R_P$-accessible to themselves; and the answer by $S$ allows us to remove all points that are $R_S$-accessible to such points that are only $R_P$-accessible to themselves. Continuing like this, the one possible answer is derived. Interestingly enough, if the upper bound 99 is relaxed (for example, 450 is taken as the upper bound), it is not the case anymore that there is a single possible answer; even other pairs $(b, l)$ with $2 \leq b \leq l \leq 99$ may then give rise to the first two lines in the conversation above!

Another interesting consideration is to determine which points are reachable from each other (see Subsection 3.1) for all the Kripke models modeling the stages in the conversation. This allows one to compute the common knowledge among $S$ and $P$ at any stage. For the first moment described in the puzzle, just before the conversation starts, the answer is given by Panti for a variant of the puzzle (where the upper bound of 99 is given as a fact but not as common knowledge) in [46]. His results imply that all points $(b, l)$ in the original Kripke model with $b + l \geq 7$ are reachable from each other. When we give the hint that indeed in the real world $b + l \geq 7$, you may deduce that at the starting point, even if the two agents have quite strong distributed knowledge, only one thing is common knowledge, namely $b + l \geq 7$.

In general, it is very difficult to establish common knowledge, especially in situations like the following, where communication is not generally known to be totally reliable.

**Example 3.7 (Byzantine generals)** Imagine two allied generals, $A$ and $B$, standing on two mountain summits, with their enemy in the valley between them[4]. It is generally known that $A$ and $B$ together can easily defeat the enemy, but if only one of them attacks, he will certainly lose the battle.

$A$ sends a messenger to $B$ with the message $b$ (= "I propose that we attack on the first day of the next month at 8 PM sharp"). It is not guaranteed, however, that the messenger will arrive. Suppose that the messenger does reach the other summit and delivers the message to $B$. Then $K_B b$ holds, and even $K_B K_A b$. Will it be a good idea to attack? Certainly not, because $A$ wants to know for certain that $B$ will attack as well, and he does not know that yet. Thus, $B$ sends the messenger back with an 'okay' message. Suppose the messenger survives again. Then $K_A K_B K_A b$ holds. Will the generals attack now? Definitely not, because $B$ does not know whether his 'okay' has arrived, so $K_B K_A K_B b$ does not hold, and common knowledge of $b$ has not yet been established.

In general, for every $n \geq 0$, one can show the following by induction. (Here, $(K_A K_B)^n$ is the obvious abbreviation for $2n$ knowledge operators $K_A$ and $K_B$ in alternation, starting with $K_A$.)

**odd rounds** After the messenger has safely brought $2n + 1$ such messages (mostly acknowledgments), $K_B(K_A K_B)^n$ is reached, but $(K_A K_B)^{n+1}$ does not hold.

**even rounds** After the messenger has safely brought $2n + 2$ such messages, one can show the following by induction: $(K_A K_B)^{n+1}$ is reached, but $K_B(K_A K_B)^{n+1}$ does not hold.

Thus, common knowledge will never be established in this way using a messenger. Moreover one can prove that in order to start a coordinated attack, common knowledge of $b$ is necessary (see [27]).

In the Byzantine generals example, the problem is, of course, that it is not guaranteed from the outset that the message $b$ will arrive at the other summit. In the next problem, we will make the circumstances a bit more favorable, and investigate whether establishing common knowledge becomes feasible (cf. [18]).

**Example 3.8** Two parties, $S$ and $R$, know that their communication channel is trustworthy, but with one small catch: when a message $m$ is sent at time $t$, it either arrives immediately, or at time $t + \epsilon$. This catch is common knowledge between $S$ and $R$. Now $S$ sends a message to $R$ at time $t_0$. Question: When will common knowledge about $m$ be established? Surprisingly, the answer is: "Never!" For an analysis in terms of distributed systems, we refer to [18]; we give a more informal explanation using Figure 4. Let the atom $s$ denote 'has been sent' and $d$ 'has been delivered'. Moreover, $e$ means 'there is a delay'. To find out whether $Cd$ will ever be established, we will, like in the muddy children example, reason from a specific situation, say $w$, in which the message was sent and delivered at time 0, without delay ($\neg e$). Moreover, we enrich our language

[4] Maybe this example from the theoretical computer scientists' folklore is not very politically correct, but you can imagine more peaceful variants in which synchronization is of vital importance, for example two robots that have to carry a heavy container together.

in the model a little by adding $s_t$ to denote that the message was sent at time $t$, and $d_t$ that it was delivered at time $t$.

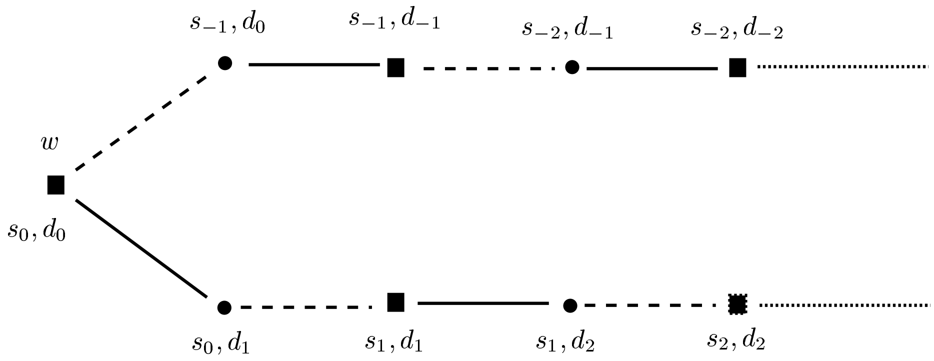


Figure 4: Message without delay

A world marked with a box denotes a world in which $e$ is false, while circles denote worlds in which $e$ holds. Accessibility of $S$ is denoted with solid arrows, that of $R$ with dashed ones. Then, at $w$ at time 0, we have $\neg Cd$, since $R$ holds it possible that there is a world where the message has a delay, for example it was sent at time zero but arrives at 1, in which world $R$ has not received the message yet: $\neg K_S \neg(e \wedge s_0 \wedge t_1)$,

How to model that $Cd$ holds at time $t$? In $w$, if $Cd$ holds, then $C(s_0 \vee s_1 \vee \ldots \vee s_t)$ should hold, but this is never the case, since, after $2t$ steps in the lower path, we have $s_{t+1}$, and hence $\neg s_t$.

**Puzzle 3.9** *What consequences would it have if the above guarantee would hold for our e-mail service?*

**Example 3.10 (Gossip)** The Dutch National Science Quiz of 1999 contained the following question, that had already circulated three decades before among mathematicians, and was solved at that time independently by Szemeredi and Tijdeman.

> Six friends each have one piece of gossip. They call each other. During each call, they share all the pieces of gossip that they know at that moment. At least how many calls are needed to bring all of them up to date with respect to the six pieces of gossip?
>
> 1. Seven.
> 2. Eight.
> 3. Nine.

During the television broadcast about the quiz on Boxing Day, none of the experts, not even the professor of mathematics in the panel, gave the correct answer. The correct answer may be found in [35], where also the minimal number of needed calls for $n$ agents with $n$ pieces of gossip is given and proved to be indeed minimal. It is clear that in the above setting, common knowledge about the pieces of gossip is not reached [13]. Thus, it would be interesting to suppose that it is possible to have $k$-conference calls (for $k \leq n$), during which subgroups of $k$ agents can communicate at the same time, hereby creating common knowledge among them about the pieces of gossip that they exchange. If $k = n$, it is clear that one $k$-conference call establishes common knowledge among all $n$ agents. An interesting question is to establish whether, if $k < n$, it is possible that the $n$ people establish common knowledge about their pieces of gossip. They are allowed to use more complex messages than conjunctions of their pieces of gossip, for example formulas that contain epistemic operators. We leave this puzzle to the reader.

# 4 Ignorance

What is a knowledge state? Is it possible to obtain a formal description of an agent's knowledge containing *exactly*, that is, at least but not more than the information conveyed by some formula $\varphi$? In other words, the case in which $\varphi$ is the agent's *only* knowledge? Characterizing an agent's exact knowledge state is important in dynamic agent systems in several ways. First of all, when the system evolves, one might wish to compare the different states of one agent: which actions (or, more specifically moves) optimally extend his knowledge? Secondly, in multi-agent systems, agent $a$ may wish to be sure that *all that b* knows is $\varphi$, and exploit the fact that $b$ does *not know more than that*. Finally, when such agents start to *exchange information*, they must be aware of principles governing their communication. Usually utterances constitute minimal information, conveying the speaker's (maximal) knowledge with respect to the relevant part of some domain (Grice's maxim of quantity, see [23]). For example, if a speaker utters the question "Do you know whether I have been rejected for the TARK conference?" The listener, having received an acceptance notice herself, may conclude that the questioner has not been accepted: for otherwise he would have received such a notice as well, and knowing the answer, he would not have asked the question. Thus, by hearing the question "$\varphi$?", the hearer concludes that indeed $\varphi$ (this example is a variant of one given by Van Benthem).

Formulas $\varphi$ representing all that the agent knows are called *honest*. In the following, as we treat the single agent case only, we may leave out the agent subscripts of the $K$-operator. For the one-agent case, some observations about only knowing and honesty are well-accepted. For instance, purely objective formulas (formulas with no occurrence of an epistemic operator) are rendered honest. If $p$ denotes 'it is Saturday' and $q$ 'it is Sunday', then waking up after a deep sleep not interrupted by an alarm clock, you can honestly claim that all you know is $(p \vee q)$. A typical example of a *dishonest* formula however, is

$\varphi = (Kp \vee Kq)$: if an agent claims to only know $\varphi$, he would know something that is stronger than $\varphi$ (i.e., either $Kp$ or $Kq$). In our example, it would not make sense to say that you only know that you know that it is Saturday or you know that it is Sunday. Similarly, one cannot honestly claim that all one knows is to know *whether* it is Saturday: $Kp \vee K\neg p$ is also dishonest.

A more sophisticated analysis of honesty generally depends on the epistemic background logic $\mathcal{S}$. The idea then is, that 'only knowing $\varphi$' corresponds to exactly knowing all $\{\psi | \varphi \vdash_{\mathcal{S}} \psi\}$. What is especially important here, is which introspective capacities we are ready to ascribe to the agent. For example, if the background logic contains the axiom of *positive* introspection $K\psi \rightarrow KK\psi$ we can infer $KKp$ if only $p$ is known by the agent. This seems innocent since the inferred knowledge is still related to the initial description $p$.

On the other hand, if we accept the axiom of *negative* introspection $\neg K\psi \rightarrow K\neg K\psi$, then we can infer knowledge concerning $q$, for example $K\neg Kq$, from only knowing $p$. This knowledge cannot be derived from only knowing $p \wedge q$, which intuitively represents *more* knowledge than only knowing $p$. This implies that we cannot just compare the knowledge states of agents. Let a knowledge state $S$ be defined as a set of formulas such that there is a $M, w$ with $M, w \models K\varphi$. Then, assuming negative introspection, $S' \subseteq S \Rightarrow S' = S$, since, if there would be a $\psi \in S$ with $\psi \notin S'$, then the agent would know (in $S'$) that he does not know $\psi$, i.e. $\neg K\psi \in S'$, but $\neg K\psi \notin S$, contradicting the assumption $S' \subseteq S$. In other words: assuming negative introspection, one agent cannot know more than another, and an agent can also not learn (since this would at the same time decrease his ignorance).

Studies on *only knowing* [25, 52] and 'all I know' [36] try to deal with this caveat. The general idea is to compare knowledge states with respect to a suitable language $\mathsf{L}^* \subseteq \mathsf{L}$. A popular restriction for $\mathcal{S}5$-agents, for example, is to take as $\mathsf{L}^*$ the objective language, the language of propositional formulas. Then, one can, for instance, say in a card-game "regarding the deal of cards, $i$ knows more than $j$". In order to formulate the syntactic approach to honesty more precisely, we need the notion of a *maximal consistent* (m.c.) set: a set $\Sigma \subseteq \mathsf{L}$ is m.c. with respect to a background logic $\mathcal{S}$ if it is consistent ($\Sigma \not\vdash_{\mathcal{S}} \bot$) and, for every formula $\sigma$: if $\Sigma \cup \{\sigma\}$ is consistent then $\sigma \in \Sigma$. The syntactic approach then judges a formula $\varphi$ honest if there exists a m.c. set $\Sigma$ with $K\varphi \in \Sigma$ and $\Sigma$ has a minimal $\mathsf{L}^*$ part: for every m.c. $\Sigma'$ with $K\varphi \in \Sigma'$ one has $\Sigma \cap \mathsf{L}^* \subseteq \Sigma' \cap \mathsf{L}^*$. Indeed, in $\mathcal{S}5$, $p \vee q$ is contained in a m.c. set with minimal propositional content, whereas for $Kp \vee Kq$, there is no such set.

In parallel to this syntactic approach comes a semantic one: how to characterise those states in which an agent knows exactly $\varphi$? Again, for $\mathcal{S}5$-agents, this is straightforward: if we have 3 atomic propositions $p, q$ and $r$, the state in which the agent knows 'nothing', or just $\top$, is any state $(M, w)$ such that *all* worlds are accessible (that is: all 8 combinations of the values of the atoms). Note that then the agent still knows all the $\mathcal{S}5$-tautologies (in particular, is still introspective) but does not know anything about objective facts: we have $(M, w) \models \neg K\varphi$, for any objective $\varphi$. Similarly, if the agent only knows $(p \vee q)$, the 'minimal' state becomes $(M', w')$ in which all but the worlds $\langle \neg p, \neg q, r \rangle$ and

$\langle \neg p, \neg q, \neg r \rangle$ are accessible.

Finally, there is a deductive component to honesty, which is also known as the Disjunction Property. Here, $\varphi$ is honest if for all $\psi_1, \psi_2 \in \mathsf{L}^*$, one has:

$$K\varphi \vdash K\psi_1 \vee K\psi_2 \quad \Rightarrow \quad (K\varphi \vdash K\psi_1 \text{ or } K\varphi \vdash K\psi_2)$$

In other words, the statement that one knows $\varphi$ must be so circumscribing, that any disjunction of knowledge can be specialized: one then has to know one of them. The Disjunction Property is often used to demonstrate that a formula under investigation is not honest: in $\mathcal{S}5$, for instance, we *do* have that $K(Kp \vee Kq) \vdash Kp \vee Kq$, but neither $K(Kp \vee Kq) \vdash Kp$, nor $K(Kp \vee Kq) \vdash Kq$.

Note that under all three approaches, the notion of honesty is related both to a subset $\mathsf{L}^*$ of $\mathsf{L}$, and to a particular modal system $\mathcal{S}$. In [32], the authors show that for any modal system $\mathcal{S}$, the three approaches are equivalent. They also suggest typical instances for the language to minimize over: two examples are $\mathsf{L}^* = \{K\psi | \psi \in \mathsf{L}\}$ and $\mathsf{L}^* = \{K\psi | \psi \in \mathsf{L}, \psi \text{ has no } K \text{ in the scope of a } \neg\}$. For $\mathcal{S}5$, the latter choice amounts to the same notion of honesty if one takes for $\mathsf{L}^*$ the objective language.

For the multi-agent case, intuition seems to be much less clear. Of course, where objective formulas are all honest in the one agent case, this property is easily convertible to formulas with no operator $K_i$, when considering honesty for agent $i$. Hence, $i$ can honestly claim to only know $K_j p \vee K_j q$, for $j \neq i$. But if $K_i$ re-occurs in the scope of $K_j$, the resulting formula $K_j p \vee K_j K_i q$ becomes *dishonest* for $i$ if $K_i$ represents (true) knowledge and $K_j$ negatively introspective knowledge. With mixed operators, in particular in the presence of negation, matters soon get fuzzy. For an overview of possible notions of honesty in a multi-agent setting, we refer to [34].

Research on 'only knowing' [25, 52] and 'all I know' [36] have largely been restricted to particular modal systems, such as $\mathcal{S}5$, $\mathcal{S}4$ and $\mathcal{K}45$. Some years ago Halpern [24] also considered other modal systems such as $\mathcal{K}, \mathcal{T}$ and $\mathcal{KD}45$. Although his approach suggests similar results for e.g., $\mathcal{KD}4$, in [32] the authors adopted a more general perspective: *given any modal system, how to characterize the minimal informational content of modal formulas.* For *multi-agent only knowing*, we only know of a (more or less) general approach by Halpern [24], putting a notion of 'possibility' to work on tree models, and, for the $\mathcal{S}5_m$ case, enriching the language with modal operators $Q_i^\zeta$, for any formula $\zeta$ and agent $i$. A thorough overview of only knowing, including studies that add an explicit operator for this notion to the language, we refer to [31].

# 5 Knowledge and games

Early game-theorists acknowledged the importance of knowledge and belief in games. A well accepted assumption is that it is common knowledge that players are rational and always maximize their utility.

**Example 5.1** Let us demonstrate the relevance of common knowledge of rationality on a small example, the game centipede (inspired by [45]). In this game, two players, say I and II, are splitting a treasure, for example a bag of marbles. They are free to choose a marble in turn, starting with player I. It is also allowed, however, that a player picks two marbles in his turn, but such a move automatically ends the game. Figure 5 illustrates an extensive form of this game for the case of 7 marbles. Picking two marbles is modelled by moving down, choosing just one marble is represented as moving right. The pay-offs for each player are summarized in the leaves of the tree (thus 1,2 means that I gets one marble and II gets two marbles). In this example, the nodes where the players may choose are labelled $1 \dots 6$.

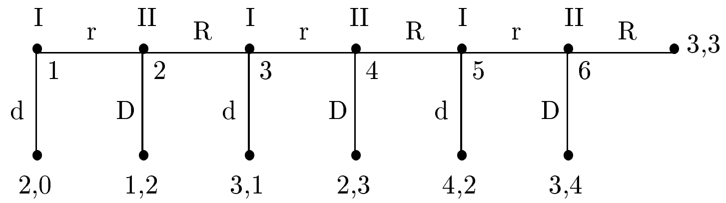


Figure 5: Six-step version of centipede

Now, let $r_1$ denote that player I is rational, and similarly $r_2$ for player II. At node 6, using that $r_2$, we can infer that II will choose Down. Since rationality of both players is common knowledge, we know that $K_I r_2$, and hence, when in node 5, player I knows that II will play 'D' in 6, and hence, since player I is rational, he prefers 4 marbles over 3 marbles and plays 'd' at node 5. Since $K_{II} K_I r_2 \wedge K_{II} r_1$, if player II were to reach node 4, he would apply the same reasoning that we just did and conclude that I will play 'd' when in 5, so II, being rational, would play 'D' in 4. Continuing this line of reasoning, and using that $K_I K_{II} K_I K_{II} K_I K_{II} r_2$, we can conclude that player I will play down at node 1 at the start!

This example illustrates the more general phenomenon of how common knowledge of rationality enables one to use backwards induction to find solutions (a Nash-equilibrium, in this case) to games, cf. [5].

In [9], Binmore distinguishes between the notions of perfect/imperfect information on the one hand, and those of complete/incomplete information on the other. A game is of *perfect information* if the rules specify that the players always know 'where they are': for games in extensive form this means that each player is free in every node to make a decision independent of that in other nodes. A game is of *complete information* if everything is known about the circumstances under which the game is played, like the probability that nature chooses a certain outcome, and who the opponent is, and how risk-averse he is.

In his thesis *Knowledge Games* [13], Hans van Ditmarsch focuses on the knowledge and ignorance of players in games with imperfect information. One of the motivations behind his work is that a player's knowledge in general improves his strategic behaviour, and hence this behaviour may be directed in finding moves that increase his knowledge. Thus, Van Ditmarsch is interested in the *dynamics* of the knowledge of the players, and tries to model the knowledge and its evolution in one and the same framework.

Van Ditmarsch introduces the notion of *knowledge games*, games in which the knowledge of the players, and the effect of their moves upon this knowledge, is described. The simplest example of a knowledge game is hexa, which provides us with a nice example of a problem which can be modelled in the spirit of *distributed systems* [18]. Here, two worlds $s$ and $t$ are accessible for player $i$ if and only if $i$'s local state in $s$ and $t$ is the same. In hexa, we have three players $(1, 2$ and $3)$ and three cards, each with a neutral side and a colored face: $r$ (red), $w$ (white) and $b$ (blue). If a player holds a card, he is the (initially only) player that knows its color. See Figure 6 for a Kripke model representing the knowledge of the players after the three cards have been dealt.

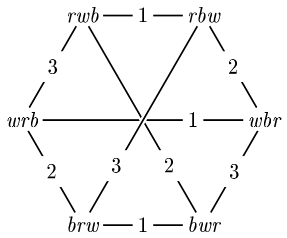


Figure 6: The model hexa for three players each holding a card

The aim of the game is to find out the distribution of the cards. Each move is comprised of a question together with its accompanying answer. We will come back to possible moves after introducing some notation. Let us abbreviate $rwb$ as the state in which player 1 holds red, 2 holds white and 3 holds blue (this distribution is denoted in the object language as $\delta_{rwb} = r_1 \wedge w_2 \wedge b_3$). Let $K_i\delta$ denote that player $i$ knows the distribution of the cards: it is an abbreviation for $\bigvee_{x,y,z\in\{r,w,b\}}^{x\neq y\neq z, x\neq z} K_i\delta_{xyz}$. The particular distribution is of course not common knowledge, but the setting of the game is: it is common knowledge which cards there are, that each player holds one card, that no card is owned by two players, etc. For instance, given the distribution $rwb$, we have that 1 knows that he has red ($K_1 r_1$), and he also knows that 2 holds blue or white ($K_1(w_2 \vee b_2)$). Finally, 1 also knows that 2 knows that either 1 has the red card, or that 3 holds it ($K_1 K_2(r_1 \vee r_3)$).

How to characterize initial situations of games like this, and, in particular, hexa? To this end, Van Ditmarsch adds the theory about this card game to the system $\mathcal{S}5_3(CE)$, where this background theory describing the initial state, just after the deal, is surprisingly simple:

**Definition 5.2**

$$\begin{array}{rcl} \mathsf{see} & := & \bigwedge_{a\in\{1,2,3\}} \bigwedge_{c\in\{r,w,b\}} (c_a \to K_a c_a) \\ \mathsf{deals} & := & \delta_{rwb} \vee \delta_{rbw} \vee \delta_{wrb} \vee \delta_{wbr} \vee \delta_{brw} \vee \delta_{bwr} \\ \mathsf{dontknowthat} & := & \bigwedge_{a\neq b\in\{1,2,3\}} \bigwedge_{c\in\{r,w,b\}} \neg K_a c_b \end{array}$$

The first property says that players know the card that they hold; the second, that the deal of cards must be one of the six ways to distribute three cards over three players; and the last property states that a player does not know the card of other players. Recall that these properties are in fact common knowledge, since $\mathcal{S}5_3(CE)$ allows us to apply Rule R3 of Definition 3.5 to them. Now, let $\mathcal{HEXA}$ be the theory $\mathcal{S}5_3(CE)$, with the properties of Definition 5.2 added to it. Then we have the following property, stating that the resulting theory yields exactly the validities of hexa:

**Theorem 5.3** $\mathcal{HEXA} \models \varphi \Leftrightarrow \mathsf{hexa} \models \varphi$

We already noted that, in knowledge games, the dynamics of epistemics is an important issue to model. Thus, we have to determine the kind of actions that are allowed in these games. To do so, suppose that 2 asks player 1: "do you have the red or the blue card?" Assuming that 1 is not allowed to lie, the following answers are possible:

($a$) Player 1 answers by saying "yes". This is not informative for player 2. However, note that player 3 now knows the distribution: he knows that there is exactly one blue card, so 1 must hold red, and hence 2 has white: $K_3\delta_{rwb}$ and hence $W_3$. Player 2 knows the latter ($K_2W_3$), i.e., 2 knows that 3 knows *what* the actual distribution $\delta$ is, although 2 does not know that 3 knows *that* it is $\delta_{rwb}$: $\neg K_2K_3\delta_{rwb}$. After 1's reply, 1 does not know that 3 can win: $(\neg K_1W_3)$, since 1 holds the distribution $rbw$ as possible, in which 1's answer would not be informative for 3.

($b_1$) Player 1 says "yes" by only showing 2 the color of his card, after which 2 can win the game by declaring the right distribution. In this case, we have common knowledge about 2's victory ($CW_2$), but of course we do not have $CK_2\delta_{rwb}$.

($b_2$) This is as ($b_1$), except that player 3 does not note that 1 shows his card to 2. Since initially, that is, before 2's question, 3 knows that 2 does not know the deal of cards ($K_3\neg W_2$), and he does not note that 1 shows his card to 2, we would have ($K_3\neg W_2 \wedge W_2$), which contradicts the property ($A3$) that knowledge is veridical. In other words, this answer of 1 would lead us out of the realm of $\mathcal{S}5_3(CE)$.

($c$) Player 1 says "yes" by putting his card publicly on the table. After this answer, it is common knowledge that 2 and 3 know the real distribution, but that 1 does not: $C(W_2 \wedge W_3 \wedge \neg W_1)$.

The only type of answer ('knowledge-action') that is allowed in the knowledge games of van Ditmarsch is that of $(b_1)$. That is, actions in which one player shows a card to another player, and this act is observed by all other players, and that is common knowledge. The result of 'applying' answer $(b_1)$ to hexa is shown in Figure 7.

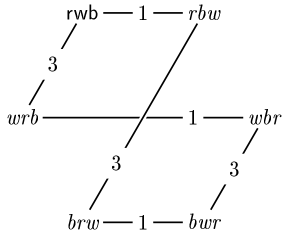


Figure 7: The model $Hexa$ after answer $(b_1)$

It is interesting that Van Ditmarsch develops a framework in which one can reason about such actions on epistemic states within one and the same framework [14, 15] that is, one does not have to 'informally' reason from Kripke model to Kripke model, but one can apply a special 'multiplication' between such models on the one hand, and action models on the other. For more details on this, we refer to [14] in this issue.

**Example 3.4** (Continued)
Let us round off this section by showing how the model of Figure 3 would evolve after the father makes his announcement $\psi$ in the case of two muddy children.

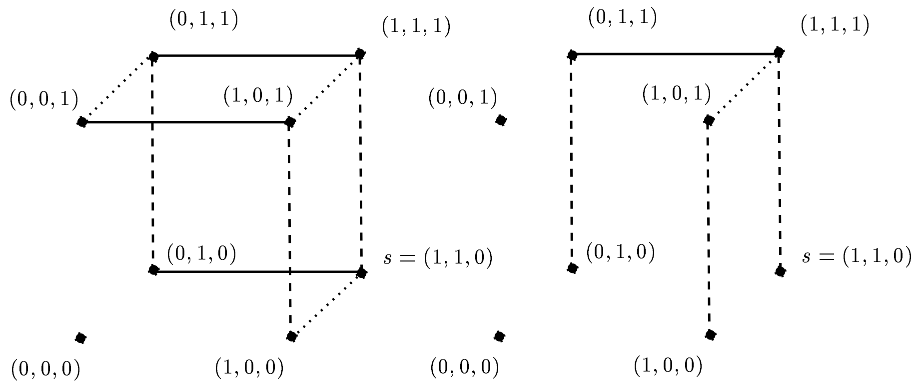


Figure 8: Father making announcements

After the first announcement of $\psi$, nobody holds the world $(0, 0, 0)$ to be compatible with his knowledge anymore, so this becomes an 'isolated world', depicted in the lefthand side model of Figure 8. In the model on the righthand side, father made his announcement $\psi$ a second time. If there would be only 1 muddy child, he would have stepped forward after the first utterance of $\psi$. Since, when $\psi$ is repeated, apparently nobody stepped forward, worlds with only one muddy child are no longer conceived possible, and hence the accessibility relations are updated accordingly. Note that in the model $M''$ obtained after two announcements, we have $M'', s \models C\psi(2)$, where $\psi(2)$ means that at least two children are muddy. But then, since in $s$, the children 1 and 2 see only one other muddy child, they know that they must be muddy themselves, and they will step forward after the second announcement (remember that we assumed that the children know the text-books on epistemic logic!). The above models would result after multiplying the Kripke model of Figure 3 with the appropriate action models.

Although Van Ditmarsch seems to be the first studying 'knowledge games' in depth, his ideas on the dynamics of epistemics are also inspired by others, especially by the work of Baltag [6] and Gerbrandy [21]. The work of Baltag is very general, in the sense that it can deal with all the examples of answers that we mentioned in the hexa, that is the answers $(a)$ to $(c)$. Gerbrandy allows for updates that need not be truthful, that is, when players are allowed to lie about their cards.

**Acknowledgements** We are thankful to Rafael Accorsi for the useful comments he gave on an earlier version of this paper.

# References

[1] M. Abadi and M. Tuttle. A semantics for a logic of authentication. In *Proceedings of the ACM Symposium on Principles of Distributed Computing*, pages 201–216, 1991.

[2] N. Agray, W. van der Hoek, and E. de Vink. On BAN logics for industrial security protocols. In B. Dunin-Keplicz and Edward Nawarecki, editors, *From Theory to Practice in Multi-Agent Systems*, number 2296 in LNAI, pages 29–36, 2002. Also at `http://link.springer.de/link/service/series/0558/papers/2296/22960029.pdf`.

[3] R.J. Aumann. Game theory. In J. Eatwell, M. Milgate, and P. Newman, editors, *Game Theory*, The New Palgrave, pages 1–54. Macmillan, 1997.

[4] R.J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

[5] R.J. Aumann and A. Brandeburger. Epistemic conditions for Nash equilibrium. *Econometrica*, (63):1161–1180, 1995.

[6] A. Baltag. A logic for supicious players. *Bulletin of Economic Research*, 54(1):1–45, 2002.

[7] P. Battagli and G. Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 52(2):149–225, 1999.

[8] M. Benereccetti, F. Giunchiglia, and L. Serafini. Model checking multiagent systems. *Journal of Logic and Computation*, 8(3):401–423, 1998.

[9] Ken Binmore. *Fun and Games. A Text on Game Theory.* D.C. Heath and Company, Lexington, MA., 1992.

[10] A. Bleeker and L. Meertens. A semantics for BAN logic, 1997. see also `http://\-dimacs.rutgers.edu/\-Workshops/\-Security/\-program2/\-program.html%`.

[11] M. Burrows, M. Abadi, and R. Needham. A logic of authentication. *ACM Transactions on Computer Systems*, 8:18–36, 1990.

[12] E. Dekel and F. Gul. Rationality and knowledge in game theory. In D. Kreps and K. Wallis, editors, *Advances in Economics and Econometrics.* Cambridge University Press, Cambridge UK, 1997.

[13] H.P. van Ditmarsch. Knowledge games. *Bulletin of Economic Research*, 53(4):249–273, 2001.

[14] H.P. van Ditmarsch. The description of game actions in cluedo. This issue, 2002.

[15] H.P. van Ditmarsch. Descriptions of game actions. accepted, 2002.

[16] S. Druiven. Knowledge development in games of imperfect information. M.Sc. thesis, in preparation, 2002.

[17] J. M. Dunn. Relevance logic and entailment. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume III. Dordrecht, 1986.

[18] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge.* MIT Press, Cambridge MA, 1995.

[19] R.F. Fagin and J.Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[20] H. Freudenthal. Problem no. 223. *Nieuw Archief voor de Wiskunde*, 3(17):152, 1969.

[21] J. Gerbrandy. *Bisimulations on Planet Kripke.* PhD thesis, University of Amsterdam, 1999.

[22] L. Gong, R. Needham, and R. Yahalom. Reasoning about belief in cryptographic protocol analysis. In *Proceedings IEEE Symposium on Research in Security and Privacy*, pages 234–248, 1990.

[23] P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Speech Acts, Syntax and Semantics III*, pages 41–58. Academic Press, New York, 1975.

[24] J.Y. Halpern. Theory of knowledge and ignorance for many agents. *Journal of Logic and Computation*, 7(1):79–108, 1997.

[25] J.Y. Halpern and Y. Moses. Towards a theory of knowledge and ignorance. In *Proc. Workshop on Non-Monotonic Reasoning, AAAI*, 1984.

[26] J.Y. Halpern and Y. Moses. A guide to the modal logics of knowledge and belief. In *Proceedings IJCAI-85*, pages 480–490, Los Angeles, CA, 1985.

[27] J.Y. Halpern and Y.O. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990.

[28] J.Y. Halpern and L.D. Zuck. A little knowledge goes a long way: Simple knowledge-based derivations and correctness proofs for a family of protocols. In *6th ACM Symposium on Principles of Distributed Computing*, pages 268–280, 1987.

[29] J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

[30] J. Hintikka. Reasoning about knowledge in philosophy: The paradigm of epistemic logic. In Joseph Halpern, editor, *Reasoning About Knowledge*, pages 63–80. Morgan Kaufmann, Los Altos, CA, 1986.

[31] W. van der Hoek, J. Jaspars, and E. Thijsse. Theories of knowledge and ignorance. Submitted to *Logic, Epistemology and the Unity of Science*, Kluwer, 2002.

[32] W. van der Hoek, J. Jaspars, and E. Thijsse. Persistence and minimality in epistemic logic. *Annals of Mathematics and Artificial Intelligence*, 27(1–4):25–47, 2000.

[33] W. van der Hoek, B. van Linder, and J.-J.Ch. Meyer. Group knowledge is not always distributed (neither is it always implicit). *Mathematical social sciences*, 38:215–240, 1999.

[34] W. van der Hoek and E. Thijsse. A general approach to multi-agent minimal knowledge: with tools and samples. *Studia Logica*, 2002. accepted.

[35] C.A.J. Hurkens. Spreading gossip efficiently. *Nieuw Archief voor de Wiskunde*, 5(1):208–210, 2000. also at `http://\-www.math.leidenuniv.nl/\-~naw/\-serie5/\-deel01/\-jun2000/`.

[36] H.J. Levesque. All I know: a study in auto-epistemic logic. *Artificial Intelligence*, 42(3):263–309, 1990.

[37] C. I. Lewis and C. H. Langford. *Symbolic Logic*. Dover Publications, New York, 1959.

[38] David Lewis. *Convention: A Philosophical Study.* Harvard U.P., Cambridge, Mass., 1969.

[39] Logic, game theory and social choice. `http//www.isdgrus.ru/LGS2/`.

[40] Logic and the foundations of game and decision theory (LOFT). `http://www.econ.ucdavis.edu/faculty/bonanno/loft.html`.

[41] W. Marrero, E.M. Clarke, and S. Jha. Model checking for security protocols. In *DIMACS Workshop on Design and Formal Verification of Security Protocols*. 1997.

[42] J. McCarthy. Formalization of two puzzles involving knowledge. In V. Lifschitz, editor, *Formalization of common sense, papers by John McCarthy*. Ablex, 1990.

[43] J.-J. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Number 41 in Cambridge Tracks in Theoretical Computer Science. Cambridge University Press, 1995.

[44] J.C. Mitchell, M. Mitchell, and U. Stern. Automated analysis of cryptographic protocols using murhpi. In *IEEE Symposium on Security and Privacy*, pages 141–153. 1997.

[45] M.J. Osborne and A. Rubinstein. *A Course in Game Theory.* MIT Press, Cambridge, Mass., 1994.

[46] G. Panti. Solution of a number theoretic problem involving knowledge. *International Journal of Foundations of Computational Science*, 2(4):419–424, 1991.

[47] R. Parikh. The logic of games and its applications. *Annals of Discrete Mathematics*, (24):111–140, 1985.

[48] F. Stulp and R. Verbrugge. A knowledge-based algorithm for the internet protocol TCP. *Bulletin of Economic Research*, 54(1):69–94, 2002.

[49] P. Syverson. The use of logic in the analysis of cryptographic protocols. In *Proceedings IEEE Symposium on Research in Security and Privacy*, 1991.

[50] Theoretical aspects of reasoning about knowledge (TARK). `http://www.tark.org`.

[51] J.F.A.K. van Benthem. Games in dynamic-epistemic logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[52] M. Vardi. A model-theoretic analysis of monotonic knowledge. In *Proceedings IJCAI85*, pages 509–512, 1985.

[53] G.H. von Wright. *An Essay in Modal Logic.* North-Holland, 1953.

[54] G. Wedel and V. Kessler. Formal semantics for authentication logics. In *Proceedings of ESORICS'96*, pages 219–241, 1996.