Direttore Responsabile: Ruggero Ferro
Iscrizione al Registro Stampa del Tribunale di Padova n. 1235 del 26.9.1990

Pubblicato con il contributo di:
Cassa di Risparmio di Padova e Rovigo

Stampa Veronese (Padova)
Laser Fotocomposizioni (Padova)

REASONING ABOUT KNOWLEDGE

Notes of ROHIT PARIKH' s Lectures

IIIrd International School for Computer Science Researchers
Acireale, June 1991, Italy

ANNA MARIA ZANABONI
Dipartimento di Scienze dell'Informazione, Università di Milano, Italy.

Contents

Reasoning about knowledge is a central issue for theories that aim to consider both the behaviour of people as social agents being part of a group, and the behaviour of artificial systems, such as computer systems. There is a strong relationship between knowledge, communication and action: what we want to study is how the action of a single agent (which is supported by his own knowledge) takes place, as a consequence of communication with other agents.

The logic of knowledge of a group of individuals is a subject of increasing interest, due on the one hand to its theoretical aspects, and on the other hand to its possible application to artificial intelligence systems, databases, distributed processing, cryptography. There are interesting combinatorial questions that arise in the case of several individuals, and do not arise in the case of one or two individuals.

The study of reasoning about knowledge motivates the following problems: how do we formalize knowledge ? how do we account for shared knowledge ? how do we account for acquisition of knowledge through communication ? how do we understand the knowledge required to perform certain actions?

Theoretical Aspects of Reasoning about Knowledge ([Hal86], [Var88], [Par90], [Mos92]) is the standard reference for discussion and comparison among researchers in this field.

1 The logic of knowledge

As will become clearer through these lecture notes, a satisfactory theoretical setting for the development of a logic of knowledge must fulfil at least three requirements:

1) *Modality*

For reasoning about knowledge in our formal system, we want to be able to represent not only facts (that is, propositions about reality), but also the knowledge each individual has about those facts, and the knowledge about the other individuals' knowledge. Modal operators can help in this respect.

2) *Non-monotonicity*

Deducing things from what cannot be proved leads to non-monotonicity. Let us consider the following dialogue between i and j:

Phase 1

i : «Guess two numbers, x and y. I tell you that x is greater than 1 and that y is greater than x. Their product is 12»

j : «I don't know what x and y are»

Phase 2

i : «Now I give you the additional information that x is odd»

j : «Now I know: x is 3».

What happened in this dialogue is that the statement «I don't know what x and y are», that was true in phase 1, does not hold any more when the additional piece of information «x is odd» is added to j's knowledge. Now x=2 can be discarded by agent j. Any logic representing the above reasoning should be non-monotonic.

3) *Vagueness*

Since there is a limit to our power of discrimination, and our language is full of vague words whose meaning can be defined only by observation, it follows that we are a linguistic community only in a loose sense. In fact, most of the times there is only a rough correspondence between what somebody says and what his interlocutor understands. The notion of meaning we want to formalize should allow for this roughness.

Epistemology—the study of knowledge—has a long tradition in philosophy, dating back to the Greeks. The first attempt to formalize reasoning about knowledge, though, can be ascribed to Von Wright [Wr51]. Hintikka [Hi62] devoted to epistemic logic the first treatment in book form.

1.1 An axiomatization

To set up the language for the logic of knowledge, we will consider a set of finitely many agents (individuals) $I = \{1, \ldots, n\}$. We assume these individuals to be interested not only in the objective reality, but also in each other's knowledge. Accordingly, our language will be the language of propositional calculus, augmented by modal operators K_i, for each $i \in I$, as follows:

a) Primitive formulae:

$P = \{p_1, \ldots, p_m, \ldots\}$ is a set of variables of the propositional calculus; they are to be interpreted as "primitive" facts.

b) Connectives:

$C = \{\neg, \wedge\} \cup \{K_i : i \in I\}$ is the set of connectives.

The K_i 's are modal operators; $K_i\varphi$ intuitively means: «agent i knows φ»

c) Well formed formulae:

WFF is the set of well formed formulae defined as follows:

1) If $p_j \in P$, then $p_j \in$ WFF;
2) If φ, $\psi \in$ WFF, then $\neg\varphi \in$ WFF and $(\varphi\wedge\psi) \in$ WFF;
3) If $\varphi \in$ WFF and $i \in I$, then $K_i(\varphi) \in$ WFF.

Let us define the following abbreviations:

1) $\varphi \vee \psi$ is equivalent to $\neg(\neg\varphi \wedge \neg\psi)$ (according to De Morgan's law)
2) $L_i(\varphi)$ is equivalent to $\neg K_i(\neg\varphi)$.
$L_i\varphi$ intuitively means: «agent i thinks φ possible»

The notion of knowledge we want to capture is conveniently axiomatized by the following set, called LK5, of axioms:

A1. All tautologies of propositional logic

A2. $K_i\varphi \wedge K_i(\varphi \rightarrow \psi) \rightarrow K_i\psi$

A3. $K_i\varphi \rightarrow \varphi$

A4. $K_i\varphi \rightarrow K_iK_i\varphi$

A5. $L_i\varphi \rightarrow K_iL_i\varphi$

R1. $\dfrac{\varphi\ ,\ \varphi \rightarrow \psi}{\psi}$ (modus ponens)

R2. $\dfrac{\varphi}{K_i\varphi}$ (necessitation).

A1 and R1 respectively are the axioms and the modus ponens rule of propositional logic. A2 states that an agent's knowledge is closed under implication, that is, if an agent i knows a formula, then he also knows all its logical consequences. A3 states that agents know only things that are true. A4 and A5 state that agents are

introspective: if an agent knows a formula, then he knows of knowing it.

There is no universal consensus on assuming the introspection axioms A4 and A5.

A2 together with R2 raise the so-called problem of "logical omniscience". They force a view of agents as "ideal knowers": adding a theorem ξ implies that $K_i\xi$ also becomes a theorem, and hence it is impossible to have $\xi \wedge \neg K_i\xi$. All agents, then, know all valid formulas and also all their logical consequences. This view may be traced back to Socrates, who, in his maieutic method, had un-educated people (like slaves) discover theorems of geometry. Nevertheless, it does not seem to be a realistic model for dealing with everyday reasoning. Indeed, even if ξ is valid, we may fail to know ξ, because of its computational intractability, or because we happen not to think of the justification for ξ, or just because we are not interested in ξ. The logical omniscience assumption has some drawbacks in data base management systems, where the available knowledge depends on computation time and space. Logical omniscience also fails in public key cryptography, where the cypher text contains the same information as the plain text, but, without knowing the key, the computation of the plain text from the cypher text is intractable. The complexity of deducing logical conclusions stems in part from the difficulty of putting together two distinct known facts—the premises—in order to draw a conclusion: computational expenses are mainly due to the binary rules of inference. This is, perhaps, the reason why we don't know as much as we should. See [Var89] for further information.

For a survey of the different approaches to the problem of logical omniscience, we refer to [Par87b], [Hal86].

Suppose we drop R2 from the above axiomatization, and add the axiom for logical omniscience:

A6. if $\vdash \xi$ then $\vdash K_i\xi$,

if $\phi \vdash \xi$ then $K_i\phi \vdash K_i\xi$,

and also add the axiom scheme:

A7. if ψ is an axiom according to A1 – A6, then so is $K_i\psi$ for each i.

Then in the new system all the old theorems are preserved, but now $\xi \wedge \neg K_i\xi$ is consistent. This is so because the new system still preserves the necessitation rule, but it states that necessitation is reasonable only for those formulae φ's which are logically true, or at least true on the whole model. That is, not necessarily from $\phi \vdash \xi$ it follows that $\phi \vdash K_i\xi$.

The above axiomatization parallels modal logic. In fact, upon reading K_i as the necessity operator, and L_i as the possibility operator, we obtain the axiom system S5 of C. Lewis [LeLa32]. This is the reason why our logic of knowledge has been called LK5. The parallel with modal logic can go further, if we observe that upon dropping axiom scheme A5, the resulting logic (called LK4) corresponds to S4. Finally, taking out also axiom scheme A4, we obtain a system (called LK) that corresponds exactly to the system T. Other possible cases (which can be located between S4 and S5) are related to temporal logic. For more details, see [Par87a].

As will be proved below, this axiomatization can be given a sound and complete semantics.

1.2 Kripke semantics

We can interpret the above logical system using models with possible worlds. Intuitively, the semantics says that beside the current state of affairs, there are other possible states of affairs (i. e., other possible worlds) for any given agent i; agents may be unable to distinguish the true world among all possible worlds. An agent is said to *know* a formula ψ if ψ is true in all the worlds that are possible for him. Nested modal operators are allowed, and, intuitively, the meaning of $K_iK_j...\varphi$ is: «agent i knows that agent j knows that ... that φ is true». In order to give a semantics to the logic of knowledge, we need a formal way of representing worlds and possibility relations (one for each agent) defined between them. Kripke structures [Kr63] are a good formal tool.

A Kripke structure M is an $(n+2)$-tuple $<S, \pi, \boldsymbol{R}_1, \dots, \boldsymbol{R}_n>$, where:

a) S is a set of states (also called *possible worlds*);

b) π is an assignment of truth values to the primitive propositions for each state $s \in S$.
$\pi(s, p_i) \in \{\text{true, false}\}$ for each state s and primitive proposition $p_i \in P$.

c) For each $i \in I$, $\boldsymbol{R}_i$ is an equivalence relation defined over S. $\boldsymbol{R}_i$ is called the *possibility* or *accessibility* relation of agent i. $(s,t) \in \boldsymbol{R}_i$ is read «t is accessible from s for agent i», or «t is i-accessible from s». $(s,t) \in \boldsymbol{R}_i$ holds iff agent i cannot distinguish the state of affairs s from the state of affairs t. In other words, if s is the actual state of the world, then agent i would consider t as a possible state of the world.

The *satisfaction* relation $\models$ for a Kripke model M is defined as follows (where $M, s \models \psi$ is read «ψ is true or satisfied in state s of model M»):

a) $M, s \models p_i$ for a primitive proposition $p_i \in P$ iff $\pi(p_i, s) = \text{true}$

b) $M, s \models \neg\varphi$ iff $M, s \not\models \varphi$

c) $M, s \models \varphi \wedge \psi$ iff $M, s \models \varphi$ and $M, s \models \psi$

d) $M, s \models K_i\varphi$ iff $M,t \models \varphi$ for all t such that $(s,t) \in \boldsymbol{R}_i$

It follows that:

e) $M, s \models L_i\varphi$ iff there exists a t such that $(s,t) \in \boldsymbol{R}_i$ and $M,t \models \varphi$

f) $M, s \models \varphi \vee \psi$ iff $M, s \models \varphi$ or $M, s \models \psi$

In particular, in light of condition d), we say that agent i *knows* fact φ in state s if φ is true at all t-accessible states—that is, all states indistinguishable from s from agent i's point of view. Moreover, the fact that each $\boldsymbol{R}_i$ is an equivalence relation ensures that everything known by i is true, and that i knows his own internal knowledge. According to condition e), we say that agent i *thinks* φ *possible* in state s if there is at least one state, say t, which is accessible from s, where φ is true.

Kripke structures can be represented by labelled graphs, whose nodes are the states in S, and two nodes s and t are connected by an arc labelled i iff $(s, t) \in \boldsymbol{R}_i$.

Despite their elegance and simplicity, there are two problems with Kripke models. First, the notion of possible world includes the psychological states of the agents. Two worlds s and t can agree on all facts, and yet they may differ only because some agent knows different things in them. This complicates the notion of possible world, and needs to be analyzed in more detail. The second problem is that, while Kripke semantics often yields a finite model property, it is true (as shown in [FHV91]) that the state of total ignorance, where nobody knows anything, can only be represented by an infinite Kripke model. For this reason, another semantics for the logic of knowledge has been proposed by [FHV88] and [MeZa85], which is equivalent to Kripke semantics, but where a possible world is represented as a tower of levels. In this model, the first level contains facts, the second contains knowledge of agents about the facts of the previous level, the k-th level contains the knowledge about the knowledge of the previous level, etc.

There is little agreement on what the possible worlds might look like. One of the problems is the so-called problem of identity. (For a wider perspective on this topic, we refer to [HuC68]). The law of substitution of equals defined as $a=b \rightarrow (P(a) \rightarrow P(b))$ is a fundamental principle governing identity. However, as Quine pointed out [Qui61], this law no longer holds in contexts involving modalities, knowledge or belief. These contexts are said to be "referentially opaque" in opposition to the "referentially transparent" ones, for which it holds. Here is an example of referentially opaque context: Ann knows that $9 > 7$. It is true that the number of planets is 9. But, Ann does not know that the number of planets is greater than 7.

1.3 Soundness and completeness

Theorem (soundness). The axiomatization of LK5 is sound.

Proof. A model M satisfies all axioms of LK5 if the $\boldsymbol{R}_i$'s are all equivalence relations. The proof follows the patterns of the usual one-agent logic of knowledge. We are only giving here the proof for axiom scheme A4 and necessitation rule R2.

a) The formula $K_i\varphi \rightarrow K_iK_i\varphi$ holds at every state s.

In fact, suppose $K_i\varphi$ holds in some M at state s. Then, for all states t such that $(s,t)\in \boldsymbol{R}_i$, φ holds at t. Now the succedent $K_iK_i\varphi$ says that for all t_1's such that $(s,t_1)\in \boldsymbol{R}_i \circ \boldsymbol{R}_i$, t_1 satisfies φ. Since $\boldsymbol{R}_i$ is transitive and $\boldsymbol{R}_i \subseteq \boldsymbol{R}_i \circ \boldsymbol{R}_i$, the t_1's are included among the t's, and the succedent is true at s.

b) Closure under necessitation.

Suppose φ is true at all states of M. Then, for any s and t, if $(s,t)\in \boldsymbol{R}_i$, then t satisfies φ. Therefore, for all s, $K_i\varphi$ holds.

It follows that all provable formulae of LK5 are true at all states of models where the $\boldsymbol{R}_i$'s are equivalence relations.

In order to prove completeness, it is sufficient to show that every consistent formula has a model of the appropriate kind. We will in fact construct a finite model.

Definition. Given a consistent formula φ, let $c(\varphi)$ denote the number of occurrences of logical symbols and atomic predicates in φ. Moreover, let F be the set of all the subformulae of φ, and W be the set of all maximal consistent

conjunctions of elements of F and their negations. Then the size of $F \leq c(\varphi)$ and so the size of W is no more than $2^{c(\varphi)}$. Let $a, b, \ldots$ denote elements of W. In our model we define π and $\boldsymbol{R}_i$ as follows:

a) For any atomic formula A, we let $\pi(a, A)$ = true iff A is a conjunct occurring in a.

b) For each individual i, let $\boldsymbol{R}_i = \{(s,t) \mid$ for all B, K_iB is in s iff K_iB is in $t\}$.

Lemma 1. Each $\boldsymbol{R}_i$ is an equivalence relation.

Proof. Obvious, from the definition.

Lemma 2. For all s in W and all ζ in F, $M,s \models \zeta$ iff B is a conjunct of s.

Proof. By induction on the complexity of ζ.

a) If ζ is atomic, the lemma holds by definition.

b) The truth-functional cases are quite easy. For example, let $\zeta = \xi \wedge \psi$.

If ζ is in F, so are ξ and ψ. Therefore, $M, s \models \xi \wedge \psi$ iff $M, s \models \xi$ and $M, s \models \psi$ iff ξ is in s and ψ is in s. Now, s contains either $\xi \wedge \psi$ or $\neg(\xi \wedge \psi)$, and is consistent. Therefore, s contains $\xi \wedge \psi$ iff it contains both ξ and ψ. Thus, $M,s \models \xi \wedge \psi$ iff s contains $\xi \wedge \psi$.

c) Let $\zeta = K_i\xi$.

If s contains $K_i\xi$, then by definition of $\boldsymbol{R}_i$, it follows that, for all t such that $(s,t)\in \boldsymbol{R}_i$, t contains $K_i\xi$, and hence, by the consistency of t, contains ξ. Therefore, for all such t's, $M,t \models \xi$, by induction hypothesis. Thus, $M,s \models K_i\xi$. On the other hand, if s does not contain $K_i\xi$, then it contains $\neg(K_i\xi)$. Arguing now by contradiction, we obtain that the set $X = \{\psi \mid K_i\psi$ is in $s\}$ is consistent with $\neg\xi$. As a matter of fact, if X is inconsistent with $\neg\xi$, then X implies ξ. Using R2, A2 and A4, we get that X implies $K_i\xi$. But X is a subset of s, and hence s is not consistent with $\neg(K_i\xi)$, a contradiction. Thus, $X \cup \{\neg\xi\}$ is consistent, and extends to some element t of W. Then $(s,t)\in \boldsymbol{R}_i$. By induction hypothesis, t satisfies $\neg\xi$, and hence s does not satisfy $K_i\xi$.

Theorem (Completeness). The axiomatization of LK5 is complete.

Proof. By Lemmas 2 and 3.

Recalling the parallel between LK and modal logic, we can make the following observations. We have just proved that the property of the $\boldsymbol{R}_i$'s of being equivalence relations guarantees soundness and completeness for LK5. It can also be proved that if we only allow the $\boldsymbol{R}_i$'s to be reflexive and transitive, they still guarantee soundness and completeness for LK4. Moreover, if the $\boldsymbol{R}_i$'s are only assumed reflexive, then they guarantee soundness and completeness for LK. The proofs for these assertions are based on two observations. First, the property of the accessibility relation that is missing for the considered axiomatic system LK_h $(h = 4, 3, 0)$ is only used to guarantee the validity of the axiom that has been dropped. Conversely, the accessibility relation that is missing for the considered axiomatic system LK_h is the only content of the axiom that has been dropped. For detailed proofs, see [Par84], [Par91b].

1.4 Decidability

The logic LK5 is decidable, i.e., there is an algorithm that, given a formula φ, tells whether it is valid or not.

Decidability is a consequence of the fact that if a formula φ is satisfiable in some model, it is also satisfiable in a model with at most 2^n states, where n is the length of φ viewed as a string of symbols.

However, deciding the validity of a formula is not an easy task. For two or more agents, any algorithm that decides the validity of a formula, requires space polynomial in the size of the formula [HaMo85]. Since it is believed that polynomial space corresponds to exponential time, the decision algorithm would require an execution time exponential in the length of the formula—which is quite unreasonable in practice. We know something more about the case of one agent: the decision procedure is NP-complete, since in this case a satisfiable formula φ` of length n can always be satisfied in a model with at most n states [Lad77]. Because of this, an open research area concerns models for resource-bounded reasoning. In [Par84] a map between LK5 and PDL^- is defined. For a survey of complexity results for logics of knowledge, see [HaV89].

1.5 Common knowledge

Common knowledge is a crucial issue for disciplines having knowledge as their subject of investigation. The first investigations on common knowledge are due to D. Lewis [Lew69] and Schiffer [Sch72]. In particular, Lewis observed that in order for something to be a convention in a group of people, this has to be common

knowledge among them. They both pointed out that co-ordinated action and proper communication require infinitely many levels of knowledge.

Although there are many approaches to common knowledge, for our purposes it is convenient to say that a group of agents has *common knowledge* of a formula φ if:

a) everyone in the group knows that φ is true (i. e., $K_i\varphi$ for all $i\in I$), and
b) everyone in the group knows that everyone knows that φ is true (i. e., $K_iK_j\varphi$ for all $i,j\in I$), and
c) everyone in the group knows that everyone knows that everyone knows that φ is true (i. e., $K_iK_jK_i\varphi$ for all $i,j\in I$),
and so on.

We can express «everybody knows φ» by writing $E\varphi$ as an abbreviation of $K_1\varphi \wedge K_2\varphi \wedge ... \wedge K_n\varphi$ for all $i,j\in I$. However, in order to convey the intuitive meaning of common knowledge, we can no longer rely on an abbreviation. We must therefore expand the language by adding a new modal operator C, where $C\varphi$ means «φ is common knowledge». We can then add the following axiom schemes to LK5, and call CK5 the resulting axiomatization :

A6. $C\varphi \rightarrow K_i\varphi$

A7. $C\varphi \rightarrow CC\varphi$

We can also define a Kripke semantics for C:

$M, s \models C\varphi$ iff $M, s \models E^k\varphi$ for each k=1, 2,...,
where $E^1\varphi = E\varphi$, and $E^{i+1}\varphi = EE^i\varphi$.

Common knowledge can also be defined in terms of the transitive closure of the union of all the $\boldsymbol{R}_i$'s. Barwise and others [Bar88] have a different view of common knowledge which we cannot go into here.

Common knowledge has become an interesting subject also in the field of economics. Aumann showed [Aum76] that if two people have the same *a priori* probabilities, and if their *a posteriori* probabilities for a given event are common knowledge, then their *a posteriori* probability must be equal. This implies that people with the same *a priori* knowledge cannot agree to disagree, and an agreement between them will always take place. Common knowledge in economics has received a complete axiomatization [Mil81]. For further information, see Section 3.

As far as distributed systems are concerned, common knowledge is an important tool in analyzing protocols for agreement. Distributedly agreed action has to rely on common knowledge: agent i sends a message to agent j, but he will know that agent j has received his message only when he receives an acknowledgement from agent j. On the other hand, agent j knows that agent i has received his acknowledgement message only when he receives an acknowledgement message from agent i, and so on. Since in systems where communication is not guaranteed common knowledge is not attainable, there are some variants of common knowledge that are attainable under more reasonable assumptions, and they are, in certain cases, indistinguishable from true common knowledge (see [HaM84], [NeT87]). For further information see Section 4.

1.6 Non-monotonicity.

Let us consider some issues about non-monotonicity, which are still under discussion among logicians (for an exhaustive discussion of what follows, see [Par84], [Par91b]. Let us go back to the dialogue mentioned in the introduction: ignorance of j in phase 1 cannot be proved in a monotonic logic, since it does not survive when a new piece of consistent information is added to the system. In fact, ignorance is due to the implicit assumption that all j knows about x and y is just what he has been told. In order to account for this fact, McCarthy proposed the following non-monotonic rule of inference to be added to LK5:

$$\frac{\text{not } (\Gamma \vdash\!- K_i A)}{\Gamma \vdash\!- \neg (K_i A)} \text{ (non-monotonicity)}$$

Intuitively, the rule says that if Γ contains all the knowledge of agent i and the formula K_iA cannot be proved from Γ, then agent i does not know A.

However, there seem to be two main difficulties with McCarthy's rule.

a) Let us consider $\Gamma = [\, K_jA \vee K_jB \,]$.

Γ does not imply K_jA, nor K_jB. By McCarthy's rule, Γ implies their negations, which, together, imply the negation of Γ. In fact, from $K_jA \vee K_jB$ we have that $\neg(\Gamma \vdash\!- K_jA)$ and $\neg(\Gamma \vdash\!- K_jB)$ both hold. Another application of McCarthy's rule yields:

$$\frac{\text{not } (\Gamma \vdash\!- K_j A)}{\Gamma \vdash\!- \neg (K_j A)} \text{ and } \frac{\text{not } (\Gamma \vdash\!- K_j B)}{\Gamma \vdash\!- \neg (K_j B)}$$

whence

$$\Gamma \vdash\!- \neg(K_jA) \wedge \neg(K_jB),$$

and, by De Morgan's rule, $\Gamma \vdash\!- \neg\Gamma$.

b) Let us consider the following more subtle argument.

Let $\Gamma = \varnothing$, and $C = (K_jA) \vee K_j(\neg K_jA)$, which is a valid formula. C doesn't imply K_jA and doesn't imply $K_j(\neg K_jA)$. Then, by McCarthy's rule, C implies both $\neg K_jA$ and $\neg K_j(\neg K_jA)$, whence C implies its own negation. Since the deduction has been obtained from the empty set, this implies the inconsistency of the empty set. Applying McCarthy's rule, we get:

$$\frac{\text{not } (\Gamma \vdash\!- K_j A)}{\Gamma \vdash\!- \neg (K_j A)} \text{ and } \frac{\text{not } (\Gamma \vdash\!- K_j (\neg K_j A))}{\Gamma \vdash\!- \neg K_j (\neg K_j A)}$$

which leads to $\Gamma \vdash\!- \neg C$, and hence $\varnothing$ is inconsistent, because C is valid.

Despite these problems, the non-monotonicity rule is intuitively a natural one. We can define some restrictions under which to apply it, in order to avoid the mentioned inconsistencies. One result is that for formulae A which have largest models, A implies B iff B has a normal proof from A using McCarthy's rule. Define a normal deduction of Γ to be one where the McCarthy rule is applied to a formula only after all subformulae of A, to which the rule might apply, have been first decided. Define also a largest model (of a specific height) of A to be one which contains a copy of any other model of A. Then we have the following completeness theorem: for those formulae A which do have a largest model M, we can deduce B from A by a normal deduction iff B is true (at the root node) of M. See [Par91b] for details.

1.7 Utility based approach to semantics

We often deal with words that can be defined only by *observation* and whose meaning is *vague*. Let us call these words *ov-words*. The judgment of an ov-word generally differs from person to person. As an example, let us consider the experiment that was done in the Acireale School: students were asked ten questions and were allowed to give a fuzzy answer (i.e., a real number $\in$ [0, 1]). All students agreed on the fact that a butterfly is an animal, but disagreement on borderline cases such as "is Sonia Gandhi an Indian?" or "is a belt an item of clothing?" still persisted. (For a detailed discussion see [Par91d]).

The fuzzy solution to the problem, proposed by Zadeh [Zad75] and others, introduces truth values between 0 and 1 and insists that there is no abrupt passage from a non-heap to a heap, but there is an infinity of possible degrees of being a heap. A bunch of grains of sand can be a heap with degree 0.8.

But the problem seems to be still there: for, we lack consensus on exactly when it is right to say that something is a heap at degree 0.8. If we can learn the truth values of "being a heap" as a fuzzy predicate, then could we speak more precisely, with a sort of fuzzy precision ? This is not easy.

However, partial social consensus on the use of many words is often adequate in order to guide actions, and people evaluate actions in terms of their utility. When ov-words are used to construct compound actions that have a higher utility value than their single components, this utility is often sturdy, and exact agreement on the use of the words is not necessary.

Let us consider the following figure:

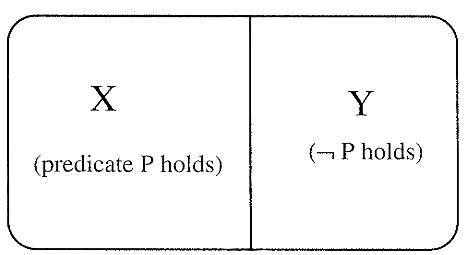

Figure 1a.
The probability space W: the probability of subspace X is p, and the probability of subspace Y is $1-p$.

Let us suppose we have a probability space W, divided into two parts, X and Y. X is the part of W on which some predicate P (for example, P = «the customer is mature») holds, and Y is the part on which it does not hold. The respective probabilities of these two parts are p and $1-p$. Suppose there are two actions, a and b, where a is: «refuse a glass of wine», and b is: «serve a glass of wine». Suppose the utility of a is 10 on X and 20 on Y. Stated otherwise, if the situation is such that P applies, then action a has a utility value of 10, but if P does not apply, then action a has a utility value of 20. Suppose the utility of action b is 10 on Y and 20 on X.

The expected utility of a on the whole space W is given by $A = 10p + 20(1-p) = 20 - 10p$, and the expected utility of b is $B = 20p + 10(1-p) = 10 + 10p$. If $p = 0.4$, then $A = 16$, and $B = 14$. The expected utility of the compound action <If P then b, else a> will instead be given by

$$20p + 20(1-p) = 20,$$

which is greater than the utility of actions a and b considered alone.

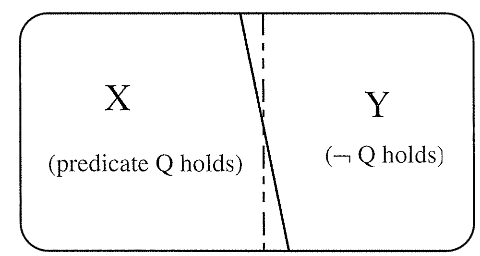

Figure 1b

Suppose now that some predicate Q (for example, Q = «the customer is over 18») roughly coincides with P, and suppose that the action associated to this partition is <if Q then b, else a>.

Suppose, for instance, that:

a) prob $(P \mid Q)$ = prob $(\neg P \mid \neg Q) = 0.9$, and
b) prob $(\neg P \mid Q)$ = prob $(P \mid \neg Q) = 0.1$.

Then the expected utility of <if Q then b, else a> will be $0.9 * 20 + 0.1 * 10 = 19$, which is greater than the probability of a and b considered alone.

In this case, if someone confuses P with Q, and does <if Q then b, else a>

instead of <if P then b, else a>, then there is not a big difference.

However, let us consider the case when the utilities for a and b are -200 and 20 (instead of 10 and 20) for a, and 20 and -200 for b (instead of 20 and 10). The action <if P then b, else a> still has an expected utility value of 20, but now the action <if Q then b, else a> has utility value -2. In this context Q is no longer a good approximation of P.

The above discussion shows that, in general, when using ov-words, we have to increase or decrease our tolerance, depending on the risks and benefits involved. How people use ov-words is not that important. What matters is that the use people make of these words, when applied to practical situations, be sufficiently similar.

For a wider set of utility functions arising from practical situations, and for a more detailed discussion, see [Par91c]. For a discussion from the viewpoint of information theory, see [KrPaN90].

2 Learning from dialogues

Dialogues are necessary in order to acquire knowledge. How do we learn from conversations ?

Considering a game between two players, we will discuss the following issues:

1) conventional proofs of the existence of knowledge correspond to optimal, complete strategies;

2) sometimes, complete strategies need to be transfinite;

3) if justified risk is preferred over absolute knowledge, then dialogues can terminate after a finite number of rounds (i. e., when, after a certain number of rounds, someone will take the risk of guessing). In this case learning can proceed faster.

We will illustrate the topic using games as examples.

Suppose we are given an arbitrary function $g : \mathbf{N} \rightarrow \mathbf{N}$. Let us consider the following game, whose players are A and B (for example, Ann and Bob): some positive number n is chosen. One of n and $g(n)$ is written on A's forehead, the other one on B's. Each player can see the other player's number, but not her/his own. Then they are repeatedly asked if they know their own number. They can answer «I don't know», or answer with the right number.

A formalization of the problem in terms of knowledge has to take into account

that each player's reasoning involves consideration about the other player's thinking and knowledge, including what the other player does not know. A's reasoning is justified if B thinks exactly as A believes B to think, and B's reasoning is justified if A thinks exactly as B believes A to think. Moreover, each player can acquire new knowledge about the situation through the dialogue.

Definition. A *stage* corresponds to a single question. A *round* consists of two stages, that correspond to one question posed to each player. The situation where A has been given number a and B has been given number b will be denoted by (a, b). The game will always start by asking A:«what is your number?», hence stages corresponding to A's turn will be odd integers, the ones corresponding to B's turn will be even integers.

2.1 Finite dialogues

Let us first consider Game 1, where $g(n) = n+1$. In this case, if the knowledge situation is (a, b), then $|a - b| = 1$. We will show by induction on the number of stages that the dialogue will always terminate with one of the players guessing her/his own number. The proof makes the assumption that each player is logically omniscient, and, moreover, that this fact is common knowledge.

Theorem 1. The following holds:

1. The player who sees the smaller number will win the game,

2. If A has the even number, then the response at the n-th stage will be «my number is $n+1$», and if A has the odd number, then the response at the $(n+1)$th stage will be «my number is $n+1$».

Proof. (By induction on $n = \min(a, b)$).

1) $n = 1$.

1a) Consider the situation $(1, 2)$, where A has number 1 and B has number 2. At the beginning of the game, A sees number 2 on B's forehead. Her number must then be 1 or 3. She doesn't know which one, so she answers «I don't know». Now B is asked. He can see on A's forehead the number 1. Since 0 is not a positive number, then B can guess «my number is 2». In conclusion, B sees the smaller number, and B guesses at stage n=2 that his number is 2 (i. e., $n+1$).

1b) Consider the situation (2, 1), where A has number 2 and B has number 1. At the beginning of the game, A sees number 1 on B's forehead. Since 0 is not a positive number, then A can guess «my number is 2» at the very first stage, i.e., at stage n=1. In conclusion, A has the even number, and A guesses at stage n=1 that her number is 2 (i. e., n+1).

2) Again, let $n = \min(a, b)$. We consider the following four cases:

<u>*Case 2a_n*</u> .

n is even, A has n, i.e., the situation is $(n, n+1) = (2k, 2k+1)$.

We are at stage n (B's turn). In this case, B sees n on A's forehead, and concludes that his own number is n-1 or n+1. If his number were n-1, then we are in case $2b_{n-1}$, and by induction hypothesis, if B's number is n-1, then A should have guessed her number at stage n-1. Since A said «I don't know my number», then B realizes that his number is not n-1, and hence it must be n+1. He then answers «my number is n+1» at stage n.

<u>*Case 2b_n*</u> .

n is odd, B has n, i. e., the situation is $(n+1, n) = (2k, 2k-1)$. We are at stage n+1 (A's turn). If n=1, then Ann would have answered at the very first stage (see 1b). If n>1, then A knows that she is either in case $2b_n$ or in case $2a_{n-1}$. In the later case, B would have guessed at stage n –1. Since he did not, case $2a_{n-1}$ is ruled out and A knows her number is n+1.

<u>*Case 2c_n*</u> .

n is even, B has n, i. e., the situation is $(n+1, n) = (2k+1, 2k)$

We are at stage n +1 (A's turn). Seeing number n on B's forehead, A knows that her number is n-1 or n+1. If it were n-1, then we are in case $2d_{n-1}$, and by induction hypothesis, if A's number is n-1, then B should have guessed his number at stage n. Since he did not, case $2d_{n-1}$ is ruled out and A knows her number is n+1.

<u>*Case 2d_n*</u> .

n is odd, A has n, i. e., the situation is $(n, n+1) = (2k-1, 2k)$

We are at stage n +1 (A's turn). If n = 1, then B would have guessed at stage 2 (see case 1a). If n>1, then an argument similar to those above applies.

The psychological arguments used in the above proof can be made rigorous by making use of suitable notions of strategies.

Definition. A Kripke model M for Game 1 consists of a space state S and two equivalence relations, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. We have $M = <S, \boldsymbol{R}_1, \boldsymbol{R}_2>$, where:

a) each state s of the state space S corresponds to a situation $s = (a, b)$ as previously defined. Stated otherwise,

$$S = \{ (m, n) \mid m, n \in \mathbf{N}^+ \text{ and } |m-n| = 1\}$$

b) state t is accessible from state s for agent $i \in \{1, 2\}$ iff when the dialogue begins, s and t are indistinguishable for agent i. Formally, $(s, t) \in \boldsymbol{R}_i$ iff $(s)_j = (t)_j$ where $j = 3 - i$ and $(s)_j$ is the j-th component of s.

Definition. A subset X of S is *i-closed* iff whenever $s \in X$ and $(s, t) \in \boldsymbol{R}_i$ then $t \in X$. X is *closed* iff it is 1-closed and 2-closed. The corresponding topologies will be denoted by T_1, T_2 and T. (A topology will be identified with the family of its closed sets). T is the intersection of T_1 and T_2 .

Definition. Given a Kripke model M, a subset X of the state space S , and a state s in X, we say that *i knows X at s* iff all $t \in S$ that are accessible from s are in X. This means that the i-closure of $\{s\}$ is contained in X. X is *common knowledge at s* iff there is a closed set Y such that $s \in X$ and $X \supseteq Y$.

Definition. A *dialogue system* DS for a Kripke model M is a map:

$$f: S \times \mathbf{N}^+ \to \{«\text{no}»\} \cup S$$

describing the players' responses at the various stages of the game. An answer «no» means «I don't know my number», whereas knowing one's number is equivalent to knowing the full state. We refer to the answer «no» as the trivial answer. All the remaining answers will be non-trivial. Formally, f satisfies the following conditions:

—for each odd number n, $f(s, n)$ (i.e., A's response at stage n) depends only on the $\boldsymbol{R}_1$ equivalence class of s and on $f(s, m)$ for $m < n$;

—for each even number n, $f(s, n)$ depends only on the $\boldsymbol{R}_2$ equivalence class of s and on $f(s, m)$ for $m < n$.

Definition. A dialogue system f is *sound* iff for all $s \in S$, if $f(s, n) \neq$ «no», then $f(s, n) = s$. Intuitively, a sound dialogue system requires that all non-trivial responses be correct.

Definition. The *index* of s, in symbols, $i_f(s)$, is the first occurrence in which a player gives a non-trivial answer. In other words, $i_f(s) =$ least n such that $f(s,n) \neq$ «no»; $i_f(s) = \infty$ if no such n exists. When it is clear from context, the subscript f will be omitted. We denote by $p(s)$ the currently acting player. Then we have: $p(s) = 1$ if $i_f(s)$ is odd, and $p(s) = 2$ if $i_f(s)$ is even.

Definition. A dialogue system f is *complete* iff $i_f(s) < \infty$ for all s.

For a sound dialogue system, $i_f(s)$ denotes the first stage when s is discovered, and $p(s)$ denotes the person who discovers it.

Definition. A sound dialogue system f is *optimal* iff for any other sound dialogue system h, we have $i_f(s) \leq i_h(s)$, for all s.

It should be noted that since a *DS* consists of a pair of strategies, one for each player, a strategy is not sound *per se*, but only in conjunction with another strategy. Now we want to characterize sound and optimal strategies. For example, the strategy of always saying «no» is sound, but not optimal.

Lemma 1. Let f be a sound *DS*. Let s and t be two distinct states, with t accessible (indistinguishable) from s for agent i. Let $i(s) = k < \infty$, and $p(s) = i$. Then $i(t) < k$ and $p(t) \neq i$.

Proof. By definition, at stage $i(s)$ player i has enough information to distinguish between states s and t. Moreover, since all previous answers associated with s have been «no», then some previous utterance associated with t must have been non-trivial. Formally, since f is sound, then $f(s, i(s)) = s \neq f(t, i(s))$. But $(s, t) \in \boldsymbol{R}_i$. Hence, there exists an integer $m < i(s)$ such that $f(s, m) \neq f(t, m)$. Since $m < i(s)$, then $f(s, m) =$ «no», and since $f(s, m) \neq f(t, m)$, then $f(t, m) \neq$ «no». This implies that $i(t) \leq m < i(s)$. Suppose now that $p(t)=i$. Then by a symmetric argument, we could also prove that $i(s) < i(t)$. But this is impossible. Hence, $p(t) \neq i$.

We can represent common knowledge as a chain. For player 1, the chain will be given by $s = s_1 \boldsymbol{R}_1 s_2 \boldsymbol{R}_2 s_3 \boldsymbol{R}_1 s_4 \ldots$, while for player 1, the chain will be of the form $s = s_1 \boldsymbol{R}_2 s_2 \boldsymbol{R}_1 s_3 \boldsymbol{R}_2 s_4 \ldots$. Elimination of a state from a chain amounts to pruning the chain after the considered state. Figure 2 describes the Kripke model for Game 1.

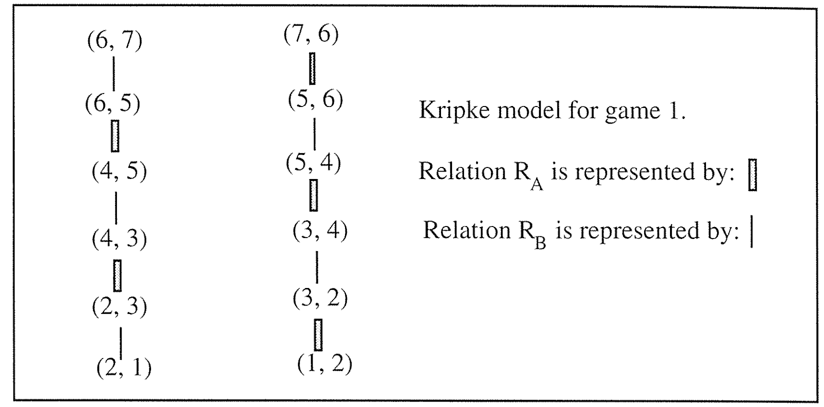

Figure 2a. Kripke model for Game 1

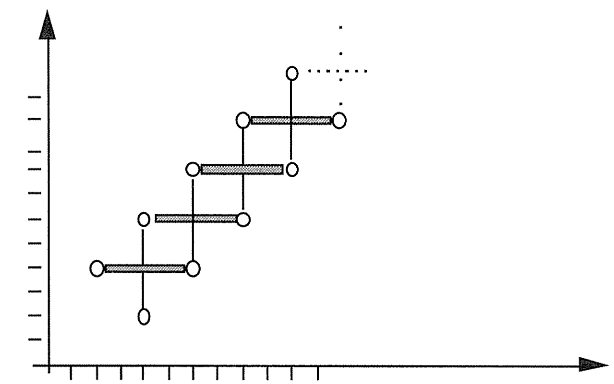

Figure 2b. The model represented on the cartesian plane

Corollary 1. Suppose $p(s) = i$, and let $s = s_1 \boldsymbol{R}_1 s_2 \boldsymbol{R}_2 s_3 \ldots\ldots \boldsymbol{R}_1 s_m$ be a chain such that no two consecutive elements are equal. Then $i(s) \geq m$.

Proof. By definition, $i(s_m) \geq 1$. Then, by induction on k, and using Lemma 1, we have that $i(s_{m-k}) \geq k + 1$. In fact, by induction hypothesis, we have $i(s_{m-k}) > i(s_{m-k+1}) = i(s_{m-(k-1)}) \geq k$. Taking $k = m-1$, we get $i(s_1) \geq m$.

Corollary 2. Suppose that there is a cyclic chain $s = s_1 \boldsymbol{R}_1 s_2 \boldsymbol{R}_2 s_3 \ldots \boldsymbol{R}_1 s_m \boldsymbol{R}_2 s_1$ with $m > 1$, any two consecutive elements being distinct. Then $i(s_m) = \infty$ for all $k \leq m$.

Proof. Assume that $p(s) = 1$. If $i(s_1) = k < \infty$, then we would get $i(s_1) > i(s_2) > \ldots > i(s_m) > i(s_1)$, which is a contradiction. On the other hand, if we assume $p(s) = 2$, then we can use the chain backwards, to get the same conclusion.

As a consequence of Corollary 2, if there are cycles in chains, no complete strategy exists. From Theorem 1 we get the soundness of the following dialogue system *f*:

a) *A*'s strategy:

- if you see an odd number $2n+1$ on *B*'s forehead, say «no» for n rounds, and then, if *B* hasn't said his number, say number «$2n+2$»;
- if you see an even number $2n$ on *B*'s forehead, say «no» for n rounds, and then, if *B* hasn't said his number, say number «$2n+1$».

b) *B*'s strategy:

- if you see an odd number $2n+1$ on *A*'s forehead, say «no» for n rounds, and then, if *A* hasn't said her number, say number «$2n+2$»;
- if you see an even number $2n$ on *A*'s forehead, say «no» for $n-1$ rounds, and then, if *A* hasn't said her number, say number «$2n+1$».

Theorem 1 amounts to saying that the above strategies yield a sound and complete dialogue system with: $i(s) = min\,(a, b)$ if *A*'s number is even, and $i(s) = max\,(a, b)$ if *A*'s number is odd.

Theorem 2. The above dialogue system (which is the same as that of Theorem 1) is optimal.

Proof. Let *f* be the given strategy, and *h* any other strategy. The proof is by cases. Suppose *A* has been given an even number, and consider the knowledge situation $s = (2n, 2n-1)$. Then, according to the strategy, we have that $i_f(s) = 2n-1$. Suppose now that in *h*, *B* is the one who first notices the state. Then we have the following chain $(2n, 2n-1)\, \boldsymbol{R}_2\, (2n, 2n+1)\, \boldsymbol{R}_1\, (2n+2, 2n+3) \ldots$, and, by Corollary 1, $i_h(s)$ cannot be finite. Suppose that in *h*, it is *A* who first discovers state *s*. In this way we have:

$$(2n, 2n-1)\, \boldsymbol{R}_1\, (2n-2, 2n-1)\, \boldsymbol{R}_2\, (2n-2, 2n-3) \ldots \boldsymbol{R}_2\, (2, 1).$$

So, by Lemma 1, $i_h(s) \geq 2n-1$. The other cases are similar.

As a second example of finite dialogues, let us consider the popular Mrs. Sum and Mr. Product puzzle [Pla89]:

Mr. Puzzle: I choose two natural numbers > 1. I will tell the sum of the numbers only to Mrs. Sum, and their product only to Mr. Product.

Mr. Puzzle proceeds to inform Mrs. Sum and Mr. Product.

1– *Mr. Product*: : «I don't know the numbers.»

2– *Mrs. Sum*: : «I knew you didn't.»

3– *Mr. Product*: : «But now I know! »

4– *Mrs. Sum*: : «So do I ! »

Now, what can be the numbers, if they both are ≤ 100 ? Let us sketch the analysis of the puzzle in terms of Kripke semantics. Our models are subsets of $W = \{ (x, y) \mid x, y$ are natural numbers and $2 \leq x \leq y \}$.

The two equivalence relations $\boldsymbol{R}_{\mathrm{Sum}}$ and $\boldsymbol{R}_{\mathrm{Product}}$ are given by:

$(x, y)\, \boldsymbol{R}_{\mathrm{Sum}}\, (z, w)$ iff $x + y = z + w$, and

$(x, y)\, \boldsymbol{R}_{\mathrm{Product}}\, (z, w)$ iff $x * y = z * w$.

According to our definitions, we say that a point (x, y) is *Sum-isolated* iff it is not $\boldsymbol{R}_{\mathrm{Sum}}$–equivalent to any other point except itself, and analogously we say that a point (x, y) is *Product-isolated* iff it is not $\boldsymbol{R}_{\mathrm{Product}}$–equivalent to any other point except itself.

The first two statements of Mrs. Sum and Mr. Product allow us to remove all Product-isolated points and all points which are $\boldsymbol{R}_{\mathrm{Sum}}$ equivalent to a Product-isolated one. From the resulting model, according to the third and fourth statements, we then remove all points which are not Product-isolated and then all points which are not Sum-isolated. Each of the remaining points can be Mr. Puzzle's choice.

An interesting related number-theoretic problem is the determination of the transitive closure of the relation $\boldsymbol{R}_{\mathrm{Sum}} \cup \boldsymbol{R}_{\mathrm{Product}}$ over *W*, or, which amounts to the same, the determination of Mrs. Sum and Mr. Product's common knowledge immediately after Mr. Puzzle's communication. This problem has recently been solved by Giovanni Panti [Pan92], a student attending the Acireale School. His result implies that the set $\{ (x, y) \in W \mid x + y \geq 7\}$ is an equivalence class of the transitive closure of $\boldsymbol{R}_{\mathrm{Sum}} \cup \boldsymbol{R}_{\mathrm{Product}}$. Equivalently, if the sum of *x* and *y* is greater than or equal to 7, then the only fact which is common knowledge is: "The sum of Mr. Puzzle's numbers is greater than or equal to 7". Figure 3 sketches the Kripke model for this puzzle. Worlds are lattice points inside the area limited by the line $x =$

2 and $x = y$. Sum-equivalence classes are segments of lines of slope -1. Product-equivalence classes are segments of hyperbolas.

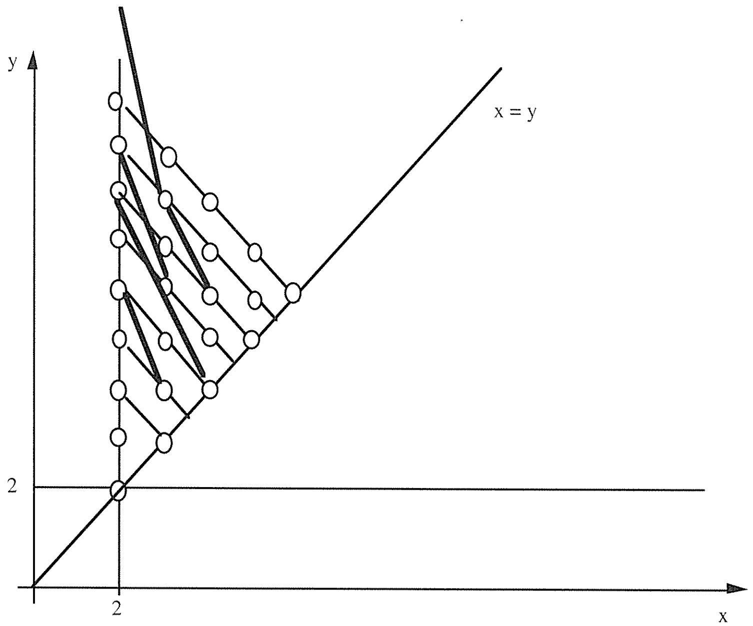

Figure 3

Kripke models for Mrs. Sum and Mr. Product represented on the cartesian plane

$\boldsymbol{R}_{\text{Sum}}$-equivalence classes are represented by lines. $\boldsymbol{R}_{\text{Product}}$-equivalence classes are represented by hyperbolas. There is no state below line $x = y$ and below line $y = 2$ and on the left of line $x = 2$, since the condition of the game is: $2 \leq x \leq y$.

2.2 *Transfinite dialogues*

Going back to the initial example mentioned at the beginning of Section 2, let us consider the following Game 2 with a more complicated function $g(n)$:

$$g(n) = \begin{cases} 1 \text{ if } n=2^k \text{ for some } k > 0 \\ n+2 \text{ if } n \text{ is odd} \\ n\text{-}2 \text{ otherwise} \end{cases}$$

Note that the numbers 2, 6, 14, ... of the form $2^k - 2$, $n \geq 2$, cannot be values of g. Graphically, the Kripke model corresponding to Game 2 can be represented as a forest composed of two trees (see Fig. 4, in which one of the trees is shown). If two states are connected by a link labelled by A, then the two values of b are the same (they are

indistinguishable for A), and if they are labelled by B, then the two values of a are the same (they are indistinguishable for B). Even though A and B may not know their own numbers, they do know which of the two subtrees they are in. The particular tree they are in is common knowledge. In this graphical representation, $g(n)$ is represented by the node just above n.

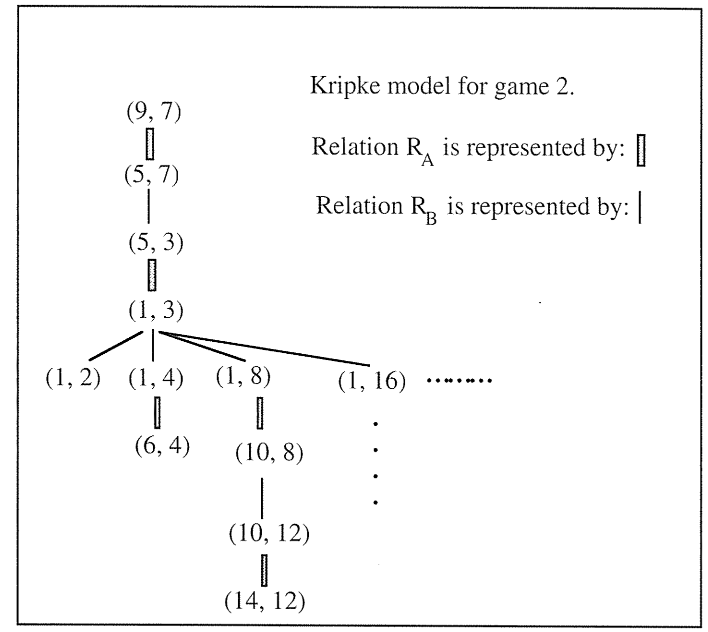

Figure 4.

Proceeding as we did for the finite dialogue case (see Theorem 1), it can be shown by induction that:

1- If the knowledge state $(n, g(n))$ or $(g(n), n)$ is below the point (1,3) in the tree and is at a distance m from a leaf, then the person whose number is of the form $g(n)$ will realize, by stage $m+2$ at most, that his/her number is $g(n)$. In particular, in state $(1, 2^k)$, A realizes by stage $2^{k-1} + 1$ that her number must be 1.

2- If the knowledge state is (1,3), B will realize, after $2^{k-1} + 1$ stages, that his number is *not* 2^k. As a consequence, after ω stages (ω being the smallest infinite ordinal), when A hasn't guessed her number yet, then B will realize that his own number is not a power of 2, and that it is therefore the number 3. That is, at the transfinite stage $\omega + 1$, one of the two players realizes his/her number.

3- If the knowledge state is above point (1,3), then A (or B) will realize his/her own number at stage $\omega+n$, where n corresponds to the steps climbed up from state (1, 3) in the Kripke structure.

Intuitively, what happens is that every time player i answers «I don't know my number», then it becomes common knowledge between the players that they are not in a state where i would have known his/her number. Such a state is always a leaf of the tree, and the trivial answer *removes* that leaf, yielding a shorter tree. Moreover, by stage ω, when A answers «I don't know my number», the entire portion of the tree below (1,3) has been pruned, and state (1, 3) becomes a leaf. At this stage, no state s is equivalent to (1, 3) for B, and, at his turn (i. e., at stage $\omega+1$), B guesses his number to be 3.

Definition. A function g is *well founded* iff there is no infinite chain $x_1, x_2, \ldots$ such that $g(x_{n+1}) = x_n$ for all n. We say that g is *finite-one* iff for all n the set $g^{-1}(n) = \{m | g(m) = n\}$ is finite.

Definition. A Kripke model M for a two agents knowledge situation consists of a space state S and two equivalence relations, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. We have $M = <S, \boldsymbol{R}_1, \boldsymbol{R}_2>$, where:

—the space S coincides with $graph(g) \cup graph(g)^{-1}$.

—state t is accessible from state s for agent $i \in \{1,2\}$ iff, when the dialogue begins, s and t are indistinguishable for agent i. Thus, $(s, t) \in \boldsymbol{R}_i$ iff $(s)_j = (t)_j$ where $j = 3 - i$ and $(s)_j$ is the jth component of s.

Definition. We define *i-closed* and *closed* sets as we did in Section 2.1. Note that topology T_1 corresponds to A's knowledge, T_2 corresponds to B's knowledge, and T corresponds to common knowledge. Also, the smallest topology T^+ containing T_1 and T_2, corresponds to shared knowledge, which is not necessarily possessed by either player individually.

Definition. Given a closed subspace X of S and a point $p \in X$, we say that p is *i-isolated in X* iff $X - \{p\}$ is *i*–closed. We also say that p is *isolated* iff it is *i*-isolated for some i. X is said to be *perfect* iff it is nonempty, closed, and has no isolated points.

Definition. Let O be the set of countable ordinals, and let M be a Kripke structure. A *transfinite dialogue system* (*TDS*) for M is a pair of maps

$$p : O \rightarrow \{1,2\}, \text{ and } f : S \times O \rightarrow \{\text{«no»}\} \cup S$$

such that for each $s \in S$ and $\alpha \in O$, $f(s, \alpha)$ depends only on the $\boldsymbol{R}_{p(\alpha)}$ equivalence class of s and on $f(s, \beta)$ for $\beta < \alpha$. Intuitively, $p(\alpha)$ denotes the person who answers at stage α, and $f(s,\alpha)$ is his/her response at stage α.

Definition. A transfinite dialogue system (p, f) is *sound* iff when $f(s, \alpha) \neq$ «no», then $f(s, \alpha) = s$, for all $s \in S$.

Definition. As for finite dialogue systems, the *index* $i_f(s)$ *of s* is the first occurrence in which a player gives a non-trivial answer. Thus, $i_f(s)$ is the smallest α such that $f(s, \alpha) \neq$ «no»; $i_f(s) = \infty$ if $f(s, \alpha)$ is always «no». By abuse of language, and consistently with our notation of Section 2.1, we write $p(s)$ instead of $p(i(s))$. We assume that A's turn occurs at all even ordinals (all limit ordinals being considered even).

It follows that for Game 2 there is no sound and complete strategy over n. But there is an optimal sound and complete strategy over the ordinals less than 2ω.

For other g, g is well founded iff the dialogue between A and B is guaranteed to terminate (with the *TDS* given above). (See [Par91a]).

2.3 *The probabilistic case*

Let us now consider the following Game 3. A number n is chosen in accordance with some probability distribution $\mu_1(n)$, for example, $\mu_1(n) = 1/(n(n+1))$. Then one player is assigned n, or $g(n)$, with probability 1/2, and the other player is assigned $g(n)$ or n, respectively. But now, instead of answering «my number is ...» only if the answer is correct, each player takes a risk of $1,000. This means that, when a player answers «my number is k», then if he is right, he will gain one dollar, and if he is wrong, he will lose 1000 dollars. μ_1 induces a probability measure μ on S given by: $\mu(n, g(n)) = \mu(g(n), n) = 1/2 * \mu_1(n)$.

Given now

a) a well-founded function g,

b) a probability distribution strictly positive on all the states,

c) a bet with fixed positive payoff for any correct guess, and fixed negative payoff for any incorrect guess,

assume that

a) each player guesses a number if and only if it is profitable for him,

b) each player answers «I don't know» when a guess is not profitable,

c) conditions a) and b) are common knowledge between the players, i. e., each one will make a guess as soon as the expected gain is greater than 0.

It follows that one of the players will always take a justified risk after a finite number of rounds, since the expected payoff will be positive for him ([Par91a]). We sketch an informal proof. Suppose g is the same as in Game 2. Suppose there are cases where there is no justification in taking the risk, and nobody takes the risk. Then there is an x of lowest rank in the tree of g such that the bet is never profitable for either side. The player who sees x knows that his number is either $g(x)$, or else it is in $X = \{y \mid g(y) = x\}$. However, by minimality of x, all these y are finitely bettable, whence it is reasonable to bet on them at some finite stage.

Therefore, as time passes, and elements of X which should have been guessed are not guessed, the set X approaches the empty set, and its probability approaches 0. Hence after some finite stage, the probability of X will be as small as needed. At this point, it will make sense to take the risk. This is a contradiction, showing that somebody is justified in taking the risk.

As an example, suppose that g is the same as in Game 2. Assume that the knowledge state is (1,3). B does not know whether his number is of the form 2^k or it is equal to 3. The probability of state (1, 3) is equal to 1/4, since: $\mu(1) = 1/2$, and the probability of $g(1)$ being assigned to B is also 1/2. On the other hand, the probability of state $(1, 2^k)$ is given by:

$$\frac{1}{2}\,\frac{1}{2^k\,(2^k+1)} = \frac{1}{2^{k+1}\,(2^k+1)}$$

The series

$$\sum_k \frac{1}{2^{k+1}\,(2^k+1)}$$

is convergent, and since there is an m (e. g., $m = 5$) such that

$$\sum_{k>m} \frac{1}{2^{k+1}\,(2^k+1)} < \frac{1}{4000}$$

it follows that after m stages in which B has ruled out numbers of the form 2^k up to 2^m, he can profitably bet on number 3 and confidently answer «my number is 3».

For our subsequent discussion, we need the following notions.

Definition. Let M be a Kripke structure, μ a probability measure on S, and $\varepsilon > 0$. A dialogue system f for the pair M,μ is said to be *ε-good* iff f satisfies the following conditions: (i) for all $s \in S$, a nontrivial answer will be eventually given at some finite stage, and (ii) if the answer given at stage n is s, then the probability that it is correct is at least $1 - \varepsilon$. In other words, there is a least natural number n such that $f(s,n) = s$, and, for this n,

$$\frac{\mu(\{s\})}{\mu(\{t \mid f(t,n) = s\})} > 1 - \varepsilon$$

Definition. A probability measure μ_1 is *computable* iff it is a computable function whose domain is the set of rational numbers.

Theorem. Let M_g be a Kripke structure arising from a well founded computable g. Suppose that μ_1 is a computable probability measure on $\mathbf{N}^+$ assigning positive probability to all $n \in \mathbf{N}^+$, and let $\varepsilon > 0$. Then, there is an ε-good computable dialogue system f for M, μ.

Proof. See [Par91a].

Theorem. g is well-founded iff for all μ, and ε there exists an ε-good dialogue system for M_g.

Proof. See [Par91a].

3 Reaching consensus through communication

How can agents in the real world reach consensus on their decisions if they rely on different pieces of information coming from different sources ? In their communication, agents exchange pieces of information that we will represent by a function f. The domain of f is the power set of W, which is the set of possible worlds, and takes values in an arbitrary set D. In many cases, agents base their decisions whether to trade or not on the expected probability of a given fixed event—

in this case, they exchange probabilities about the event (f is a probability distribution). Or, as in the stock market case, agents exchange just their decisions (f has values «yes», «no»: agents just declare if they are willing to trade. For a detailed discussion of this case, see [Kra90], [PaKr90]).

Assuming agents to behave in a rational way and to communicate according to some fair protocol, how does learning take place, and how is consensus reached ?

Depending on the properties of f (see below for definitions), three main results will be discussed, namely:

1- If the function f satisfies the union consistency (sure thing) property, then consensus on the values of f can be reached in case of two participants in the conversation;

2- If the function f is weakly convex, then consensus can be reached in case of at most three participants;

3 - If the function f is (strongly) convex, then consensus can be reached in case of any number n of participants, where the communication is pairwise and the protocol of communication is fair.

Moreover, in these cases 1-3, consensus is always reached even in the absence of common knowledge ([PaKr90]). This result is very important, since in most real cases and applications we would like to use consensus (which is a "low level" kind of knowledge) without demanding common knowledge (which is a very "high level" kind of knowledge).

Definition. Let $I = \{1, \dots n\}$ be a set of individual agents. Let W be a state space and $\boldsymbol{R}_1 \dots \boldsymbol{R}_n$ be the accessibility relations for agents $1,\dots n$. Let $P_1, \dots, P_n$ be n partitions of W, each P_i corresponding to individual $i \in I$. Two states of W, p and q, belong to the same element P_i iff $(p, q) \in \boldsymbol{R}_i$ (where the P_i's are equivalence classes of W). P^+ denotes the coarsest common refinement (join) of the P_i's. P^- denotes the meet of the P_i's. We then have, for all $i \in I$: $P^- \le P_i \le P^+$.

Definition. A subset X of W is said to be *i-closed* iff $p \in X$ and $(p, q) \in \boldsymbol{R}_i$ imply $q \in X$. In other words, an i–closed set is a union of elements of P_i).
X is *weakly closed* iff it is a union of equivalence classes of P^+. An i-closed set is also weakly closed, but the converse is not always true.

Definition. A *protocol Pr* is a pair of functions specifying, at each discrete point in time, the sender and the receiver of the communication. Formally, $Pr = (s(t), r(t))$, where $s(t)$ and $r(t)$ have the set of natural numbers **N** as their domain, and the set $\mathbf{N}^+$ of positive integers as their codomain. As usual, t stands for time.

Definition. A protocol Pr is *fair* iff every agent is a recipient and a sender infinitely many times, and each agent receives information from every other agent (possibly in an indirect way) infinitely many times. This means that everybody participating in the communication will have access to the pieces of information being exchanged.

More formally, consider a directed graph E whose nodes are agents $i \in I$ and such that there is an edge (i, j) connecting node i to node j iff for infinitely many times t we have $s(t) = i$ and $r(t) = j$. Then the protocol is fair if the graph is strongly connected: from every node there is a path to every other node of the graph.

An example of fair protocol is "round-robin", in which the first participant sends a piece of information to the second one, who in turn sends a piece of information to the third participant, and so on, until the last agent sends a piece of information to the first one, and the cycle is repeated indefinitely. Formally, the round robin protocol can be defined in the following way.

Let $I = \{0, \dots n-1\}$ be the set of participants in the protocol. Then $s(0) = 0$, $r(0) = 1$, and for $t \ge 1$, $s(t) = r(t-1) = t \ mod \ n$. We stipulate that the pieces of information that are exchanged during the conversation are represented by a function f whose domain is the power set of W, and taking values in an arbitrary set D.

Definition. We denote by $m(t)$ the message sent at time t, and by $C(x, i, t)$ the set of possible states of the world for agent i at time t, given that the real state is $x \in W$. We define $m(t)$ and $C(x, i, t)$ by induction on t as follows:

$$m(t) = f(C(x, s(t), t)),$$

$$C(x, i, 0) = P_i(x) = \{ y \mid (x,y) \in \boldsymbol{R}_i \},$$

$$C(x, i, t+1) = \begin{cases} C(x, i, t) \cap NewInf & \text{if } i = r(t), \text{ where } NewInf = \{y \mid m(y,t) = m(x,t)\} \\ C(x, i, t) & \text{otherwise} \end{cases}$$

This definition also gives an updating procedure for each agent's state of mind: when new pieces of information are given, the recipient agent can eliminate possibilities from his state of mind. We denote by $C(x,i,\infty)$ the limit value of $C(x,i,t)$, whenever it exists. If there is an x such that $p = f(C(x,i,\infty))$ then we will call p a *possible limiting value* for i.

3.1 Union consistency property

Definition. The function f satisfies the *sure thing principle* (also known as the *union consistency property*) iff for all disjoint subsets X and Y of W with $f(X) = f(Y)$, we have $f(X \cup Y) = f(X) = f(Y)$.

Theorem. If f satisfies *union consistency*, then consensus is reached if there are only two agents.
Proof. See [GePo82], [Cav83], [Bac85].

As an example, let us consider the situation where W is the set of possible results of an experiment, and there are two agents interested in computing the probability p of some fixed event E, in order to make their decisions. Each agent receives information about the experiment by being told one of the elements of his partition P_i of W. We assume that the P_i's are common knowledge among agents, and agents always receive true information (i. e., if the real state of the world is x, then $x \in P_i(x)$ for all i). The function f is a conditional probability on a set of events (the codomain D of f is the set of real numbers). Agents communicate values of f.

Without additional information, every agent would have the same value $p(E)$. Let us instead consider the following procedure:

a) one agent (the sender) learns that the result x is in $P_i(x)$, computes a new (a-posteriori) probability of E, $p(E|P_i(x))$, and sends it to the other agent;
b) the other agent (the receiver) in turn computes his new probability eliminating from his set of possible worlds all the worlds in which the first agent would have sent a different value, and sends the new computed value to the first agent,

and so on.

In this case, since the conditional probability function satisfies the sure thing principle, and the a posteriori probabilities are common knowledge, then they must coincide. This can be interpreted by saying that like-minded agents cannot agree to disagree. For the original setting of the problem see [Aum76]; for further discussion, see [GePo82], and for its reformulation using the sure thing principle see [Cav83] and [Bac85].

The same result holds with more than two agents, in *broadcast modality*, in the

sense that everyone can hear any agent communicating his own value of f. But the result no longer holds if participants in the conversation are more than two and the communication of f is not public. (For three agents, a counter-example can be found in [PaKr90]). Most real situations correspond to this case, where agents communicate by pairwise private interaction. As we will see in 3.2, the property of weak convexity of f guarantees consensus if the agents are at most three. We will also show in 3.3 that the stronger property of (strong) convexity of f guarantees consensus for any number of agents.

3.2 Weak convexity

Definition. The function f is *weakly convex* iff for all closed disjoint subsets X and Y of W, there are two numbers, $a, b \geq 0$, such that $a + b = 1$ and $f(X \cup Y) = a * f(X) + b * f(Y)$.

This definition implies that, if $X_1, \ldots, X_k$ are closed and pairwise disjoint sets, then there exist reals $a_1, \ldots, a_k$ such that: $a_1 + \ldots + a_k = 1$ and $f(X_1 \cup \ldots \cup X_k) = a_1 * f(X_1) + \ldots + a_k * f(X_k)$. Clearly, weak convexity implies union consistency.

Theorem. Suppose f is weakly convex, the protocol Pr is fair, and three participants communicate values of f according to Pr. Then consensus on the value of f must be reached.
Proof. See [PaKr90].

This result does not hold for $n > 3$ agents. A counter example (where $n = 4$) can be found in [PaKa90].

3.3 Convexity

Definition. A function f is *convex* iff for all closed disjoint subsets X and Y of W, there are $a, b > 0$, such that $a + b = 1$ and $f(X \cup Y) = a * f(X) + b * f(Y)$.

Convexity implies the union consistency property (the proof is the same as for weak convexity), and it also implies weak convexity. In fact, convexity is a stronger property: everything that lies in the open interval between $f(X)$ and $f(Y)$ also lies in the closed interval.

Theorem. If f is convex, consensus is reached for any number n of participants in the conversation, provided that communication is pairwise, and the protocol of communication is fair. More precisely, there is an integer t_0 such that for all $x \in P_i$, $i \in I$, and all $t, t' > t_0$, we must have $C(x, i, t) = C(x, i, t')$. In particular, for all x and i the limiting value $f(C(x, i, \infty))$ exists. Moreover, if f is convex and the protocol is fair, then this limiting value does not depend on i.

Proof. For all x and i, $C(x, i, t)$ is a decreasing function of t. Further, $C(x, i, t)$ is a nonempty union of P^+-equivalence classes. Since P^+ is finite, $f(C(x, i, t))$ is eventually constant. Let $t(x, i)$ be the least t such that for every $t' > t$, the value of $C(x, i, t')$ is constant. Then $t(x, i)$ depends only on the P^+–equivalence classes of x. Since P^+ has a finite number of equivalence classes, and there are finitely many i, then there is a t_0 such that $C(x, i, t')$ is constant for all $t' > t_0$ regardless of x and i. Clearly, t_0 depends on the choice of protocol $Pr = (s(t), r(t))$. It follows that for all $t' > t_0$ the value $f(C(x, s(t), t))$ depends only on x and $s(t)$. Assume for simplicity that 3 agents communicate according to a round robin protocol. Let:

$p_1 < p_2 < \ldots < p_k$ be the possible limiting values for participant 1,
$q_1 < q_2 < \ldots < q_r$ be the possible limiting values for participant 2,
$r_1 < r_2 < \ldots < r_m$ be the possible limiting values for participant 3.

We have to show that for all x, the limiting values generated by x do not depend on i. Let p_i be a possible limiting value for agent 1, and $E(p_i)$ be the set of all x such that $f(C(x, 1, \infty)) = p_i$. Then $E(p_i)$ is a disjoint union of sets $C(x, 1, \infty)$, namely, those sets for which $C(x, 1, \infty) = p_i$. If $y \in C(x, 1, \infty)$, then y is compatible with all the information received by agent 1 when the real world is x. This implies that y and x are equally informative for agent 1, whence $C(y, 1, \infty) = C(x, 1, \infty)$. Therefore, by the union consistency principle, which is in turn implied by convexity, we have $f(E(p_i)) = p_i$. Let us say that p_i and q_j are compatible iff there exists an x such that $f(C(x, 1, \infty)) = p_i$ and $f(C(x, 2, \infty)) = q_j$. This means that, for such an x, agent 1 sends the value p_i to agent 2. After many iterations of the protocol we have

$$C(x, 2, t+1) = C(x, 2, t) \cap \{y \mid f(C(y, 1, t)) = p_i\} = C(x, 2, t) = C(x, 2, \infty).$$

Therefore, the set $\{y \mid f(C(y, 1, t)) = p_i\}$ must contain $C(x, 2, \infty)$. Thus $E(p_i)$ contains $C(x,2,\infty)$. This implies that $E(p_i)$ is a union of sets of the form $C(x, 2, \infty)$ whence, by the assumed convexity of f, p_i must be an average of those q_j which are compatible with it. Let p_1 be the smallest of the p_i's. The set of values q_j compatible with p_1 must contain at least one element, less than or equal to p_1. It follows that $q_1 \le p_1$. By symmetry, $r_1 \le q_1$ and $p_1 \le r_1$. Then $p_1 = q_1 = r_1$. But only q_1 can be compatible with p_1, since the remaining q_j are strictly greater than p_1, and the average of q_1 with any other q_j is strictly greater than q_1. Similarly, only r_1 can be compatible with q_1, and only p_1 can be compatible with r_1. Thus, after time t_0, either all or none of agents 1, 2, 3 will be sending p_1. Thus in particular, p_2 can be compatible only with q_j for $j \ge 2$, q_2 can be compatible only with r_h for $h \ge 2$, and r_2 can be compatible only with p_i for $i \ge 2$. Repeating the argument, we see that $p_2 = q_2 = r_2$. By induction we get $p_d = q_d = r_d$ for all d. Now, our argument is independent of n and of the particular protocol. In fact, the only property we used is that we can find a chain $t_1 < t_2 < \ldots < t_u$ with all t_i greater than t_0 and such that (a) $s(t_1) = 1$, (b) the sender at time t_{j+1} is the recipient at time t_j, and (c) the chain passes through all participants, finally returning to 1. Now observe that conditions (a)-(c) are automatically satisfied by any fair protocol.

References

[Aum76] R. J. AUMANN, Agreeing to disagree, *Annals of Statistics*, **4** (1976) 1236-1239.

[Bac85] M. BACHARACH, Some extensions of a claim of Aumann in an axiomatic model of knowledge, *Journal of Economic Theory*, **37** (1985) 167–190.

[Bar88] J. BARWISE, Three views of common knowledge, Theoretical Aspects of Reasoning about Knowledge (edited by Vardi), Morgan-Kaufmann, 1988, 365–380.

[Cav83] J. CAVE, Learning to agree, *Economics Letters*, **12** (1983) 147–152.

[FHV91] R. FAGIN, J. HALPERN, M. VARDI, A model-theoretic analysis of knowledge, *Journal of ACM*, 1991, 382–428.

[Goe86] K. GOEDEL, Collected Works; S. Feferman et al. (Editors), Oxford University Press, 1986.

[Hal86] J. Y. HALPERN, Reasoning about Knowledge: an Overview, In: *Theoretical Aspects of Reasoning about Knowledge* (edited by Halpern), Morgan-Kaufmann, 1986, 1–18.

[HaM85] J. Y. HALPERN, Y. O. MOSES, A guide to the modal logics of knowledge and belief: preliminary report, *Proc. of the 9th IJCAI*, 1985.

[HaM84] J. Y. HALPERN, Y. O. MOSES, Knowledge and common knowledge in a distributed environment, *Proc. of the 3rd ACM Conference on Principle of Distributed Computing*, 1984, 50–61.

[HaV89] J. Y. HALPERN, M. VARDI, The complexity of reasoning about knowledge and time, *Journal of Computer and System Sciences*, **38** (1989).

[Hi62] J. HINTIKKA, Knowledge and Belief, Cornell University Press, Ithaca, NY, 1962.

[HoU79] J. E. HOPCROFT, J. D. ULLMAN, Introduction to Automata Theory, Languages and Computation, Addison-Wesley, 1979.

[HuC68] G. E. HUGHES, M. J. CRESSWELL, An Introduction to Modal Logic; Methuen, London, 1968.

[Kra90] P. KRASUCKI, Reaching consensus on decisions, Theoretical Aspects of Reasoning about Knowledge (edited by Parikh), Morgan-Kaufmann, 1990, 141–150.

[KrPaN90] P. KRASUCKI, R. PARIKH, G. NDJATOU, Probabilistic knowledge and probabilistic common knowledge, *Methodologies for Intelligent Systems*, **5** (1990), (edited by Ras, Zemankova, Emrich), Elsevier, pp. 1–8.

[Lad77] R. E. LADNER, The computational complexity of provability in systems of modal propositional logic, *SIAM Journal of Computing*, **6** (1977) 3.

[LeLa32] C. I. LEWIS, C. H. LANGFORD, Symbolic Logic, Dover, Mineola, NY, 1932.

[Lew69] D. LEWIS, Convention, a Philosophical Study, Harvard University Press, 1969.

[Mak66] D. MAKINSON, On some completeness theorems in modal logic, *Zeit. f. math. Logik*, **12** (1966).

[MeZa85] J. MERTENS, S. ZAMIR, Formulation of bayesian analysis in games with incomplete information, *International Journal of Game Theory*, **14** (1985).

[Mi81] P. MILGROM, An axiomatic characterization of common knowledge, *Econometrica*, **49** (1981)

[Mos92] Y. MOSES (ed.), Theoretical Aspects of Reasoning about Knowledge, Morgan-Kaufmann, 1992

[NeT87] G. NEIGER, S. TOUEG, Substituting for real time and common knowledge in asynchronous distributed systems, *6th ACM Symposyum on Distributed Computing*, 1987, 281–293.

[PaKr90] R. PARIKH, P. KRASUCKI, Communication, consensus and knowledge, *Journal of Economic Theory*, **52** (1990) 178–189.

[Pan92] G. PANTI, Solution of a Number Theoretic Problem Involving Knowledge, to appear in: *International Journal Found. Comput. Sci.*, 1992

[Par84] R. PARIKH, Logics of knowledge, games and dynamic knowledge, In: Proc. Foundations of Software Theory and Theoretical Computer Science, *Lecture Notes in Computer Science*, **181** Springer-Verlag, 1984.

[Par87a] R. PARIKH, Modal Logic, In: *Encyclopedia of Artificial Intelligence*, Wiley & Sons N.Y., 1987.

[Par87b] R. PARIKH, Knowledge and the problem of logical omniscience, *ISMIS-87*, (edited by Ras, Zemankova), North-Holland, 1987.

[Par90] R. PARIKH, Recent issues in reasoning about knowledge, Theoretical Aspects of Reasoning about Knowledge (edited by Parikh), Morgan-Kaufmann, 1990, 3–10.

[Par91a] R. PARIKH, Finite and infinite dialogues, *Proc. Workshop on Logic and Computer Science*, MSRI (edited by Y. Moskovakis), Springler-Verlag , 1991.

[Par91b] R. PARIKH, Monotonic and non-monotonic logics of knowledge, *Fundamenta Informaticae*, **15** (1991).

[Par91c] R. PARIKH, A utility based approach to semantics, Research Report, 1991.

[Par91d] R. PARIKH, A test for fuzzy logic, *SIGACT News*, **22** (1991), 49–50.

[Pla89] J. A. PLAZA, Logics of public communications, *Proc. Fourth International Symposium on Methodologies for Intelligent Systems*, Oak Ridge National Laboratory, 1989, 201–216.

[Qui61] W. V. QUINE, Reference and Modality, In: *From a Logical Point of View*, Harper and Row, New York, 1961.

[Sch72] S. SCHIFFER, Meaning, Oxford University Press, 1972.

[Var88] M. VARDI (ed.), Theoretical Aspects of Reasoning about Knowledge, Morgan-Kaufmann, 1988

[Var89] M. VARDI, On the complexity of epistemic reasoning, *4th IEEE-LICS Symposium*, 1989.

[Wr51] G. H. VON WRIGHT, An Essay in Modal Logic, North-Holland, Amsterdam, 1951.

[Zad75] L. ZADEH, Fuzzy logic and approximate reasoning, *Synthese* (special issue on the Logic Semantics of Vagueness), **30** (1975) 407–428.