REASONING ABOUT KNOWLEDGE: AN OVERVIEW

Joseph Y. Halpern
IBM Research Laboratory
Almaden, CA 95193

Abstract: In this overview paper, I will attempt to identify and describe some of the common threads that tie together work in reasoning about knowledge in such diverse fields as philosophy, economics, linguistics, artificial intelligence, and theoretical computer science. I will briefly discuss some of the more recent work, particularly in computer science, and suggest some lines for future research.

1. Introduction

Although *epistemology*, the study of knowledge, has a long and honorable tradition in philosophy, starting with the Greeks, the idea of a formal logical analysis of reasoning about knowledge is somewhat more recent, going back to at least von Wright ([Wr]). The first book-length treatment of epistemic logic is Hintikka's seminal work, *Knowledge and Belief* [Hi1]. The 1960's saw a flourishing of interest in this area in the philosophy community. Axioms for knowledge were suggested, attacked, and defended. Models for the various axiomatizations were proposed, mainly in terms of possible-worlds semantics, and then again attacked and defended (see, for example, [Ge,Len,BP]).

More recently, reasoning about knowledge has found applications in such diverse fields as economics, linguistics, artificial intelligence, and computer science. While researchers in these areas have tended to look to philosophy for their initial inspiration, it has also been the case that their more pragmatic concerns, which often centered around more computational issues such as the difficulty of computing knowledge, have not been treated in the philosophical literature. The commonality of concerns of researchers in all these areas has been quite remarkable. Unfortunately, lack of communication between researchers in the various fields, while perhaps not as remarkable, has also been rather noticeable.

In this overview paper, I will attempt to identify and describe some of the common threads that tie together research in reasoning about knowledge in all the areas mentioned above. I will also briefly discuss some of the more recent work, particularly in computer science, and suggest some lines for future research. This should by no means be viewed as a comprehensive survey. The topics covered clearly reflect my own biases.

2. The "classical" model

We'll begin by reviewing the "classical" model for knowledge and belief (now over 25 years old!), the so-called *possible-worlds* model. The intuitive idea here is that besides the true state of affairs, there are a number of other possible states of affairs, or possible worlds. Some of these possible worlds may be indistinguishable to an agent from the true world. An agent is then said to *know* a fact φ if φ is true in all the worlds he thinks possible. For example, an agent may think that two states of the world are possible: in one it is sunny in London, while in the other it is raining in London. However, in both these states it is sunny in San Francisco. Thus, this agent knows that it is sunny in San Francisco, but does not know whether it is sunny in London.

The philosophical literature has tended to concentrate on the one-agent case, in order to emphasize the properties of knowledge. However, many applications of interest involve multiple agents. Then it becomes important to consider not only what an agent knows about "nature", but also what he knows about what the other agents know and don't know. It should be clear that this kind of reasoning is crucial in bargaining and economic decision making. As we shall see, it is also relevant in analyzing protocols in distributed computing systems (in this context, of course, the "agents" are the processors in the system). Such

reasoning can get very complicated. Most people quickly lose the thread of such nested sentences as "Dean doesn't know whether Nixon knows that Dean knows that Nixon knows about the Watergate break-in"; (this example comes from [ClM], where the point is investigated further). But this is precisely the type of reasoning that goes on in proving lower bounds for certain distributed protocols (cf. [HM,DM]).

In order to formalize this situation, we first need a language. The language I'll consider here is a propositional modal logic for m agents. Starting with primitive propositions p, q, r, ..., more complicated formulas are formed by closing off under negation, conjunction, and the modal operators K_1, ..., K_m. Thus, if φ and ψ are formulas, then so are $\sim\varphi$, $\varphi \wedge \psi$, and $K_i\varphi$, $i = 1, ..., m$. This last formula is read "agent i knows φ". The K_i's are called modal operators; hence the name modal logic. We could also consider a first-order modal logic that allows quantification, but the propositional case is somewhat simpler and has all the ingredients we need for our discussion.

Kripke structures [Kr] provide a useful formal tool for giving semantics to this language. A Kripke structure M is a tuple $(S, \pi, \mathscr{P}_1, ..., \mathscr{P}_m)$, where S is a set of *states* or *possible worlds*, π is an assignment of truth values to the primitive propositions for each state $s \in S$ (so that $\pi(s,p) \in \{\text{true}, \text{false}\}$ for each state s and primitive proposition p), and $\mathscr{P}_i$ is an equivalence relation on S for $i = 1, ..., m$ (recall that an equivalence relation is a binary relation which is reflexive, symmetric, and transitive). $\mathscr{P}_i$ is agent i's *possibility relation*. Intuitively, $(s,t) \in \mathscr{P}_i$ if agent i cannot distinguish state s from state t (so that if s is the actual state of the world, agent i would consider t a possible state of the world).

We now define a relation $\models$, where $M, s \models \varphi$ is read "φ is true, or *satisfied*, in state s of model M":

$M, s \models p$ for a primitive proposition p if $\pi(p, s) = \text{true}$

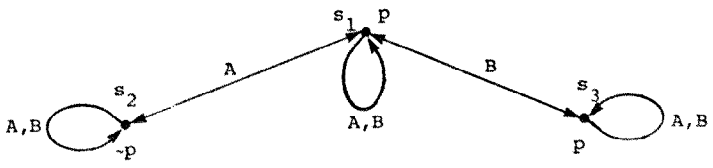
$M, s \models \sim\varphi$ if $M, s \not\models \varphi$

$M, s \models \varphi \wedge \psi$ if $M, s \models \varphi$ and $M, s \models \psi$

$M, s \models K_i\varphi$ if $M, t \models \varphi$ for all t such that $(s,t) \in \mathscr{P}_i$.

The last clause is designed to capture the intuition that agent i knows φ exactly if φ is true in all the worlds that i thinks are possible.

These ideas are perhaps best illustrated by an example. One advantage of Kripke structures is that we can easily represent them as labelled directed graphs, where the nodes are exactly the states in S and two nodes s and t are connected by an edge labelled i exactly if $(s,t) \in \mathscr{P}_i$. Consider the graph below, where $S = \{s_1, s_2, s_3\}$ and there are two agents, Alice and Bob. Assume for simplicity there are only one primitive proposition in the language, say p. We can think of p as standing for "it is sunny in San Francisco". Thus, in state s_1, it is sunny



in San Francisco, but Alice doesn't know it (since she considers both s_1 and s_2 possible). On the other hand, she does know that Bob knows whether or not it is sunny in San Francisco (since in both states she considers possible, Bob knows the weather at that state). Bob knows that it is sunny in San Francisco, but he doesn't know that Alice doesn't know this fact (since he considers s_3 possible, and in s_3 Alice does know it!). Formally, we have

$$M, s_1 \models p \wedge \sim K_A p \wedge K_A(K_B p \vee K_B \sim p) \wedge K_B p \wedge \sim K_B \sim K_A p.$$

Note that in both s_1 and s_3, the primitive proposition p gets the same truth value. One might think that in some sense, therefore, s_1 and s_3 are the same, and one of them can be eliminated. This is not true! A state is not completely characterized by the truth values that the primitive propositions get there. The possibility relation is also crucial. For example, at s_1, Alice doesn't know p, while at s_3 she does. Even with only one primitive proposition, there are non-trivial models with infinitely many states.

The notion of knowledge defined here has a number of interesting technical properties. It can be shown that if a formula φ is satisfiable in some model, it is satisfiable in a model with at most 2^n states, where n is the length of φ viewed as a string of symbols. From this result, it follows that the logic is decidable: there is an algorithm that, given a formula φ, can tell whether or not it is *valid* (i.e., true in every state of every model). However, deciding validity is not easy. Any algorithm that does so requires space polynomial in the size of the input formula ([HM2]). Since we believe polynomial space corresponds to exponential time, this would suggest it would also require time exponential in the size of the formula, a quite unreasonable requirement in practice. The one-agent case is somewhat simpler. As shown by Ladner [La], a satisfiable formula in the one-agent case can in fact always be satisfied in a model with at most n states. As a consequence, the decision procedure in the one-agent case in NP-complete. However, we also believe that NP-complete problems require exponential time in practice, so even the one-agent case is quite difficult. (See [HU] for an introduction to complexity theory.)

The notion of knowledge we have been considering can be completely characterized by the following *sound* and *complete* axiom system, due to Hintikka ([Hi1]); i.e., all the axioms are valid and every valid formula can be proved from these axioms.

A1. All instances of propositional tautologies.

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. $\sim K_i\varphi \Rightarrow K_i \sim K_i\varphi$

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

R2. $\dfrac{\varphi}{K_i\varphi}$

A1 and R1, of course, are holdovers from propositional logic. A2 says that an agent's knowledge is closed under implication, A3 says that an agent only knows things that are true. This is the axiom that is usually taken to distinguish *knowledge* from *belief*. You cannot know a fact that is false, although you may believe it. A4 and A5 are axioms of introspection. Intuitively, they say that an agent is introspective: he can look at his knowledge base and will know what he knows and doesn't know. There are numerous papers in the philosophical literature discussing the appropriateness of these axioms (cf. [Len]). Philosophers have tended to reject both of the introspection axioms for various reasons.

The validity of A3, A4, and A5 is due to the fact that we have taken the $\mathscr{P}_i$'s to be equivalence relations. In a precise sense, A3 follows from the fact that $\mathscr{P}_i$ is reflexive, A4 from the fact that it is transitive, and A5 from the fact that it is symmetric and transitive. By modifying the properties of the $\mathscr{P}_i$ relations, we can get notions of knowledge that satisfy different axioms. For example, by taking $\mathscr{P}_i$ to be reflexive and transitive, but not necessarily symmetric, we retain A3 and A4, but lose A5; similar modifications give us a notion that corresponds to belief, and does not satisfy A3. (See [HM2] for a survey of these issues, as well as a review of the standard techniques of modal logic which give completeness proofs in all these cases.)

However, the possible-worlds approach seems to commit us to A2 and R2. This forces us to a view of our agents as "ideal knowers", ones that know all valid formulas as well as all logical consequences of their knowledge. This certainly doesn't seem to be a realistic model for human agents (although it might perhaps be acceptable as a first approximation). Nor does it seem to even be an adequate model for a knowledge base which is bounded in terms of the computation time and space in memory that it can use. We discuss some approaches to this problem of *logical omniscience* in Section 5 below.

3. A concrete interpretation: distributed systems

While it is not clear whether or not the model presented above is appropriate for human reasoning, it can capture quite well much of the reasoning that goes on in analyzing distributed systems. Indeed, the distributed systems point of view allows us to give quite a concrete interpretation to states in a Kripke structure.

A distributed system consists of a collection of processors, say $1, \ldots, m$, connected by a communication network. The processors communicate with each other over the links in the network. Each processor is a state machine, which at all times is in some state. This state is a function of its initial state, the messages it has received, and possibly some internal events (such as the ticking of a clock). The *global state* of the system is just a description of each processor's state. We can associate a Kripke structure with a distributed system by taking the states in the structure to be all the possible global states of the system over time. The $\mathscr{P}_i$ relations are defined by $(s,t) \in \mathscr{P}_i$ if processor i has the same state in global states s and t. Note that this definition makes $\mathscr{P}_i$ an equivalence relation. The primitive propositions

in this setting would be statements like "the value of processor i's local variable x is 0" or "processor j's current state is σ".

This model, or slight variants of it, has appeared in many recent papers in distributed systems (cf. [HM1,PR,HF,ChM,DM,FI,LR,FV2]). Note that in this model, knowledge is an "external" notion. We don't imagine a processor scratching its head wondering whether or not it knows a certain fact φ. Rather, a programmer reasoning about a particular protocol would say, from the outside, that the processor knows φ because in all global states consistent with its current state (intuitively, all the global states that the processor could be in, for all it knows) φ is true.

This notion of knowledge is information based, and does *not* take into account, for example, the difficulty involved in computing knowledge. Nor could a processor necessarily answer questions based on its knowledge, with respect to this definition of knowledge. So on what basis can we even view this as knowledge? When trying to prove properties such as lower bounds on the number of rounds required to complete a given protocol, the kinds of arguments that one often hears have the form "We can't stop after only three rounds, because processor 1 might not know that processor 2 knows that processor 3 is faulty." Now this informal use of the word "know" is exactly captured by the definition above. Let φ state that processor 2 knows that processor 3 is faulty. Then processor 1 doesn't know φ exactly if there is a global state of the system that it cannot distinguish from the actual state where φ does not hold; i.e., where processor 2 doesn't know that processor 3 is faulty.

It is interesting to note that essentially the identical notion of knowledge was developed independently by Rosenschein and his coworkers (cf. [Ro,RK]) and used for describing and analyzing situated automata in AI applications.

4. Some variants of the "classical" model

While we can give states in a Kripke structure a clear interpretation when we are considering distributed systems, their interpretation is not so clear when we try to model human reasoning, even assuming we are dealing with "ideal knowers". Recall that a state cannot be characterized simply by the primitive propositions that are true in that state. A better characterization of a state would consist of the primitive propositions that are true there, together with the set of states that each agent cannot distinguish from that state (this, after all, is the information carried by the edges of the graph corresponding to the Kripke structure). But this is circular: we are characterizing a state in terms of other states!

The circularity can be broken by taking the knowledge structures approach described in [FHV]. A *knowledge structure* is constructed inductively as follows. A depth 0 world is just a truth assignment to the primitive propositions. A depth 1 world essentially consists of a set of depth 0 worlds for each agent, intuitively, all those depth 0 worlds that the agent thinks possible; a depth 2 world essentially consists of a set of depth 1 worlds for each agent, etc. (There are also some consistency conditions that these worlds must satisfy, but we omit them here.) The limit of this construction is a knowledge structure.

There is a very close relationship between knowledge structures and Kripke structures. Every state in a Kripke structure corresponds to a knowledge structure where the same formulas are true, and every knowledge structure corresponds to a state in some Kripke structure. Thus the same axioms characterize knowledge structures and Kripke structures. By modifying the consistency conditions that worlds must satisfy, we can get knowledge structures that do not necessarily satisfy A3, A4, and A5 ([FHV,FV1,Va1]). However, just as for Kripke structures, we do seem to be committed to A2 and R2.

Interestingly, a parallel development is apparent in the economics literature on knowledge. The first economics paper to give a formal model for knowledge is that of Aumann [Au]; his model is essentially identical to a Kripke structure, but with the added feature of a probability measure on the set of states.[1] Then Mertens and Zamir constructed a model that is essentially a knowledge structure, carrying along the probability at each level (cf. [MZ]). By including probability in the picture in this way, we can reason about the probability an agent assigns to another agent assigning a certainly probability that certain propositions are true, and so on. Such an analysis is critical in, for example, game-theoretic arguments.

5. The problem of logical omniscience

As we mentioned above, all the models that have been described so far assume that agents are ideal knowers, who know all valid formulas and all logical consequences of their knowledge. Clearly this is not a realistic view of human agents; it is also an inappropriate model for a number of other applications. What features an appropriate model should have may depend on the application.

One approach that has frequently been suggested is the syntactic approach: what an agent knows is simply represented by a set of formulas (cf. [Eb,MoH]). Of course, this set need not be constrained to be closed under logical consequence or to contain all instances of a given axiom scheme. While this approach does allow us to define a notion of knowledge that doesn't suffer from the logical omniscience problem, it is a notion that is extremely difficult to analyze. If knowledge is represented by an arbitrary set of formulas, we have no principles to guide a knowledge-based analysis. A somewhat more sophisticated approach is taken by Konolige ([Ko]), who considers starting with a set of base facts, and then closing off under a (possibly incomplete) set of deduction rules.

One semantic approach that has been taken is to augment the standard possible worlds by "impossible" worlds, where the customary rules of logic do not hold (cf. [Cr1,Ra,RB]). However, these impossible worlds have not been very well motivated (although see [Hi2] for motivation for one of these models). More recently, Levesque ([Lev2]) has given a more

---

1 Actually, instead of considering equivalence relations, Aumann viewed the $\mathscr{P}_i$'s as *partitions* of S, where a partition is a set of disjoint subsets whose union is all of S. But a partition is just another way of looking at an equivalence relation. More formally, given a partition $\mathscr{P}$ of S, we can construct the equivalence relation $\mathscr{P}'$ where $(s,t) \in \mathscr{P}'$ iff s and t are in the same subset of $\mathscr{P}$. Conversely, given an equivalence relation $\mathscr{P}'$, we can construct the partition $\mathscr{P}$ whose subsets consist of the equivalence classes of $\mathscr{P}'$.

intuitively plausible semantic approach. He distinguishes between *implicit* knowledge and *explicit* knowledge, where explicit knowledge consists of those facts that you are explicitly aware of, while implicit knowledge consists, intuitively, of all the logical consequences of explicit knowledge. Of course, an agent may not be aware of all his implicit knowledge. It is implicit knowledge that satisfies all the axioms we discussed above. Levesque considers a possible-worlds model where, in a given state, a primitive proposition may be either true, false, both (so that the state is inconsistent), or neither.[2] He then develops a logic of implicit and explicit belief based on this approach. (Although Levesque considers belief rather than knowledge, his results can easily be extended to the case of knowledge.) Explicit belief implies implicit belief, but not conversely. It turns out that $B\varphi \Rightarrow B\psi$ holds in Levesque's logic for propositional formulas φ and ψ exactly if φ *entails* ψ in *relevance logic*,[3] where B is the modal operator for explicit belief (Levesque only considers the one-agent case, so the B is not subscripted). Levesque's logic avoids the logical omniscience problem, in that it is not the case that if φ is a tautology, then $B\varphi$ holds. However, an agent in Levesque's logic is still a perfect reasoner as far as relevance logic may be concerned. While it is not clear that humans are any better at relevance logic than propositional logic, there may still be some interesting applications to Levesque's ideas. For example, Levesque has shown that for an interesting subclass of formulas in his logic (namely, those of the form $B\varphi \Rightarrow B\psi$, where φ and ψ are propositional formulas in conjunctive normal form), the validity problem can be decided in polynomial time. Such results indicate that this may be a useful logic for a knowledge base to use. An efficient algorithm could be written which would allow a knowledge base to answer questions based on what it knows, with respect to this notion of knowledge.

Levesque's polynomial time results do not seem to extend once we allow meta-reasoning (i.e., an agent reasoning about his own knowledge) or consider several agents. We remark, though, that by extending Levesque's ideas, Patel-Schneider ([Pa]) and Lakemeyer ([Lak]) have designed semantics for a first-order logic of knowledge which has an interesting decidable fragment.

Fagin and Halpern have taken a slightly different approach to this issue ([FH]). Their *logic of general awareness* is essentially a mixture of syntax and semantics. It starts with a standard Kripke structure, and adds to each state a set of formulas that the agent is "aware" of at that state. Implicit knowledge is defined just as knowledge was before. Explicit knowledge consists of implicit knowledge plus awareness. Thus an agent explicitly knows φ if φ is true in all the worlds an agent considers possible and φ is in the awareness set for

---

2 Levesque calls his states *situations*, and they are essentially the situations of *situation semantics* (cf. [BP]). For further details on the situation semantics approach to modelling knowledge and belief, the interested reader is encouraged to consult [BP].

3 Relevance logic is a weakening of propositional logic that uses a notion of entailment rather than implication. It was motivated by a desire to avoid some of the well-known paradoxes of implication in propositional logic, such as the fact that $(\varphi \Rightarrow \psi) \vee (\psi \Rightarrow \varphi)$ is a tautology for any formulas φ and ψ. See [AB] for further discussion.

that agent. We can give these models a concrete interpretation along the lines of the distributed system interpretation described above by imagining that each processor is running some algorithm to compute what it knows, given the information it has received. The awareness set is just the set of formulas it can figure out the truth of, using this algorithm, in some prespecified time or space bound.

Let me mention one final approach to the problem of logical omniscience, this one inspired by work of Montague. Montague gives a possible-worlds semantics to epistemic logic (in that formulas are still associated with sets of possible worlds), but knowledge is not modelled as a relation between possible worlds; i.e., there is no $\mathscr{P}_i$ relation ([Mon]; see also [Va2]). By doing this we lose the intuition that an agent knows φ exactly if φ is true in all worlds that agent considers possible. But this approach does provide a handle to getting around the logical omniscience problem. One deficiency in Montague's semantics is that, while agents need not know all logical consequences of their knowledge, they are unable to distinguish between logically equivalent formulas. To solve this, Thomason went a step further by supplying a model-theoretic semantics that does not use possible worlds [Th] (see also [Cr2]).

6. Common knowledge

A persistent theme in almost every discipline that has considered knowledge at all is the study of *common knowledge* and its cousins, such as *mutual belief*. There are many approaches to defining common knowledge (see [Ba] for a discussion). For our purposes, we can take it to be the case that when a group has common knowledge of φ, then not only does everyone in the group know that φ is true, but everyone knows that everyone knows, everyone knows that everyone knows that everyone knows, etc. Note that if we take $E\varphi$ to represent "everyone knows φ" and $C\varphi$ to represent "φ is common knowledge", then it is straightforward to give semantics to these formulas in Kripke structures:

$M,s \models E\varphi$ if $M,t \models \varphi$ for all t such that $(s,t) \in \mathscr{P}_1 \cup \ldots \cup \mathscr{P}_m$

(so that $M,s \models E\varphi$ iff $M,s \models K_1\varphi \wedge \ldots \wedge K_m\varphi$)

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

This key notion was first studied by David Lewis, in the context of conventions ([Lew]). Lewis points out that in order for something to be a convention, it must be common knowledge among the members of the group. It also arises in discourse understanding. If Ann asks Bob "Have you ever seen the movie playing at the Roxy tonight?", then in order for this question to be interpreted appropriately, not only must Ann and Bob know what movie is playing tonight, but Ann must know that Bob knows, Bob must know that Ann knows that Bob knows, etc. (this is discussed in great detail in [ClM], although see [PC] for a slightly dissenting view).

Interest in common knowledge in the economics community was inspired by Aumann's seminal result [Au]. Aumann showed that if two people have the same priors, and their posteriors for a given event are common knowledge, then these posteriors must be equal. This result says that people with the same priors *cannot agree to disagree*. Since then, common

knowledge has received a great deal of attention in the economics literature, with issues such as complete axiomatizations ([Mi]) and the number of rounds of communication information required before the posteriors for an event become common knowledge ([GP]) being examined. (Further results and references can be found in [MS,Ca,TW].)

The questions that have arisen in distributed systems work on knowledge are surprisingly similar to those of economics. Again the key observation is that agreement implies common knowledge of the agreement, so that common knowledge becomes an important tool in analyzing protocols for agreement. To help illustrate this point, consider the *coordinated attack problem* from the distributed systems folklore ([Gr]):[4]

> Two divisions of an army are camped on two hilltops overlooking a common valley. In the valley awaits the enemy. It is clear that if both divisions attack the enemy simultaneously they will win the battle, whereas if only one division attacks it will be defeated. The divisions do not initially have plans for launching an attack on the enemy, and the commanding general of the first division wishes to coordinate a simultaneous attack (at some time the next day). Neither general will decide to attack unless he is sure that the other will attack with him. The generals can only communicate by means of a messenger. Normally, it takes the messenger one hour to get from one encampment to the other. However, it is possible that he will get lost in the dark or, worse yet, be captured by the enemy. Fortunately, on this particular night, everything goes smoothly. How long will it take them to coordinate an attack?

Suppose the messenger sent by general A makes it to general B with a message saying "Let's attack at dawn". Will general B attack? Of course not, since general A does not know he got the message, and thus may not attack. So general B sends the messenger back with an acknowledgement. Suppose the messenger makes it. Will general A attack? No, because now general B does not know he got the message, so he thinks general A may think that he (B) didn't get the original message, and thus not attack. So A sends the messenger back with an acknowledgement. But of course, this is not enough either. I will leave it to the reader to convince himself that no amount of acknowledgements sent back and forth will ever guarantee agreement. Note that this is true even if the messenger succeeds in delivering the message every time. All that is required in this reasoning is the *possibility* that the messenger doesn't succeed.

This rather convoluted reasoning can easily be expressed in terms of knowledge. Each time the messenger makes a transit, the depth of the generals' knowledge increases by one, so it goes from "B knows" to "A knows that B knows" to "A knows that B knows that A

---

4 The following discussion is taken from [HM1].

knows" (that the attack is to be held at dawn), and so on. However, they never attain common knowledge that the attack is to be held at dawn by this protocol. Indeed, it can be shown (see [HM1]), that in any system where communication is not guaranteed, common knowledge is not attainable. Since it can also be shown that in a precise sense agreement implies common knowledge, it follows as a corollary that the generals cannot agree to a coordinated attack *in any run of any protocol.* (Of course, this does not rule out the possibility of a probabilistic notion of agreement, so that with high probability they both attack.)

Interestingly, Halpern and Moses show that not only is common knowledge not attainable in systems where communication is not guaranteed, it is also not attainable in systems where communication *is* guaranteed, as long as there is some uncertainty in message delivery time. Thus, in practical distributed systems, common knowledge is not attainable. This remark holds for systems of communicating humans as well as processors. What is going on here? After all, we often do reach agreement (or seem to!). Common knowledge is attainable in "idealized" models of reality where we assume, for example, events can be guaranteed to happen simultaneously. Even if we cannot always make this assumption in practice, it turns out that there are some variants of common knowledge that are attainable under more reasonable assumptions, and these variants are indistinguishable in certain cases from true common knowledge (see [HM1]). Such variants may prove a useful tool for specifying and analyzing different assumptions on communication in distributed systems (see also [FI] for further discussion along these lines).

7. Knowledge, communication, and action

Implicit in much of the previous discussion has been the strong relationship between knowledge, communication, and action. Indeed, much of the motivation for studying knowledge by researchers in all areas has been that of understanding the knowledge required to perform certain actions, and how that knowledge can be acquired through communication. I will just briefly touch on some of the more recent trends in this area.

Early work of McCarthy and Hayes [McH] argued that a planning program needs to explicitly reason about its ability to perform an action. Moore [Mo] took this one step further by emphasizing the crucial relationship between knowledge and action. Knowledge is necessary to perform actions, and new knowledge is gained as a result of performing actions. Moore went on to construct a logic with possible-worlds semantics that allowed explicit reasoning about knowledge and action, and then considered the problem of automatically generating deductions within the logic. This work has recently been extended by Morgenstern [Mor]; she views "know" as a syntactic predicate on formulas rather than a modal operator.

Another issue that has received a lot of attention recently is the relationship between knowledge and communication. Levesque considered this from the point of view of a knowledge base that could interact with its domain via *TELL* and *ASK* operations ([Lev1]). He showed, somewhat surprisingly, that the result of *TELL*ing a knowledge base an arbitrary sentence in a first-order logic of knowledge is always equivalent to the result of *TELL*ing it

a purely first-order sentence (i.e. one without any occurrences of K). It is worth remarking here that it is crucial to Levesque's result that there is only one knowledge base, i.e. one agent, in the picture.

Characterizing the states of knowledge that result after communication is also surprisingly subtle. One might think, for example, that after telling someone a fact p they will know p (at least, if it is common knowledge that the teller is honest). But this is not true. For example, consider the sentence "p is true but you don't know it". When told to agent i, this would be represented as $p \wedge \sim K_i p$. Now this sentence might be perfectly true when it is said. But after i is told this fact, it is not the case that $K_i(p \wedge \sim K_i p)$ holds. In fact, this latter formula is provably inconsistent! It is the case, though, that i knows that $p \wedge \sim K_i p$ was true before, although it is no longer true now.

Even if we do not allow formulas that refer to knowledge, consider the difficulty of characterizing the knowledge of an agent Alice that has been told only one fact: the primitive proposition p. Intuitively, all she knows is p. Since we are assuming ideal agents, Alice also knows all the logical consequence of p. But is this all she knows? Suppose q is another primitive proposition. Surely Alice doesn't know q, i.e. $\sim K_A q$ holds. But we assume Alice can do perfect introspection, so that she knows about her lack of knowledge of q. Thus $K_A \sim K_A q$ holds. But this means that even if "all Alice knows is p", then she also knows $\sim K_A q$, which is surely not a logical consequence of p! The situation can get even more complicated if we let Bob into the picture. For then Alice knows that Bob doesn't know that Alice knows q (how can he, since in fact she doesn't know q, and Bob does not know false facts). And knowing that Bob can also do perfect introspection, Alice knows that Bob knows this fact; i.e., $K_A K_B \sim K_B K_A p$ holds! Thus, despite her limited knowledge, Alice knows a nontrivial fact about Bob's knowledge (see [FHV,HM3] for further discussion of these points). Part of the difficulty here is due to *negative introspection*, i.e., the fact that one has knowledge about one's own lack of knowledge. If we remove this feature from our model (i.e., discard axiom A5), things become much easier (cf. [Va1]).

A related issue is characterizing what states of knowledge are attainable as a function of the communication medium. [HM1] already shows that while common knowledge is a perfectly consistent state of knowledge, it is not attainable if communication is not guaranteed or even if it is guaranteed and there is uncertainty about the message transmission time (see [ChM,FI] for related results). [FV2] provides a complete characterization of the states of knowledge of the current world attainable in a synchronous system where communication is not guaranteed. However, many open questions still remain in this area. Typical examples include: characterizing asynchronous systems, proving bounds on the number of messages required to attain certain attainable states of knowledge (cf. the results of [GP]), and dealing with knowledge about the past and future.

Most of the work discussed above has implicitly or explicitly assumed that the messages received are consistent. The situation gets much more complicated if messages may be

inconsistent. This quickly leads into a whole complex of issues involving belief revision and reasoning in the presence of inconsistency. Although I won't attempt to open this can of worms here, these are issues that must eventually be considered in designing a knowledge base, for example, since there is always the possibility of getting inconsistent information from a user (see [Be] for some further discussion on this topic).

8. Areas for Further Research

I will conclude now with what I consider to be a number of hard problems for further research. Although they are listed separately below, it should be clear that there is a lot of overlap among them. Many of these problems have already been mentioned in this overview. Of course, these are only *some* of the important problems in the field; the list is by no means an exhaustive.

1. *Models for resource-bounded reasoning.* Find models appropriate to capture resource-bounded reasoners, particularly models that incorporate time. Another desideratum for such models is that they be powerful and flexible enough to capture various theories of learning. There have actually been a number of papers that have considered logics of knowledge and time (cf. [Sa,Leh,LR,HV]). Fagin and Halpern [FH] show that interesting properties of bounded reasoning can be captured by letting the awareness set vary over time. However, no attempt has been made to systematically study and model situations where such change occurs.
2. *Models that take into account the cost of acquiring knowledge.* These may very well end up being special cases of models that model resource-bounded reasoning, but the idea of cost of knowledge is so important that I include it as a special case here. Cost of acquiring knowledge is particularly relevant in such areas as cryptography and economics. Goldwasser, Micali, and Rackoff's work on *knolwedge complexity* ([GMR]), which attempts to quantify the amount of information released during an interaction in terms of computational complexity theory, may provide the right approach to tackling this problem. The idea here is that you gain extra information as a result of an interaction exactly if you can perform a computation using your limited (polynomial-time) resources that you could not perform before the interaction.
3. *Logics of knowledge with easy decision procedures.* Find well-motivated sublanguages that are simple enough to allow a computer to completely carry out all reasoning in that sublanguage, yet powerful enough to capture many features of interest. (See [Lev2,Pa,Lak] for some steps in this direction.) These logics should allow meta-reasoning (reasoning about one's own knowledge) and reasoning about the knowledge of other agents if at all possible.
4. *Knowledge, action, and communication.* Find good models for the interaction between knowledge and action, especially when the knowledge is partial. Characterize the states of knowledge attainable under different assumptions on the communication medium.
5. *Knowledge and cryptography.* Relate some of these models to work going on in understanding cryptographic protocols. The idea is to formally analyze, for example, *public key*

cryptosystems such as that of [RSA]. Here all the information required to break the code is known in principle (since the encoding key is published publicly, say in a telephone book), but the code is still hard to break since (we believe) it is very hard to compute the decoding key from the encoding key. Some preliminary steps have already been taken (see, for example, [Me,GMR]), but much work remains to be done. Note that the cost of computing knowledge becomes particularly important here, as well as analyzing probabilistic knowledge. Being able to break the code with high probability is essentially as good as knowing how to break the code.

The problems listed above are far from completely specified. Indeed, a large part of the difficulty of the problem lies in finding a precise formulation of it amenable to attack. And, at least as important, is the problem of finding a formulation that is intuitively natural, and can really be used to clarify the analysis of important problems in areas such as economics, linguistics, or distributed computation. Further experience with using knowledge to specify and reason about distributed systems and AI systems should help to enhance our understanding as well as guiding further research. Indeed, by focussing in on such applications we can often get a much better grasp of how to handle more general problems. I am optimistic that major advances can be made on a number of fronts in the near future.

Because reasoning about knowledge is an area that lies at the intersection of a number of fields, any advances made could have wide-spread impact. But in order for this to be true, researchers in the relevant communities need to be more aware of each other's work. I hope that this paper will help open the lines of communication.

Acknowledgements: I'd like to thank Ron Fagin, Hector Levesque, Yoram Moses, and Moshe Vardi for their useful comments on an earlier draft of this paper.

References

[AB] A. R. Anderson and N.D. Belnap, *Entailment, the Logic of Relevance and Necessity*, Princeton University Press, 1975.

[Au] R.J. Aumann, Agreeing to disagree, *Annals of Statistics*, 1976, pp. 1236-1239.

[Ba] J. Barwise, Modelling shared understanding, unpublished manuscript, 1985.

[BP] J. Barwise and J. Perry, *Situations and Attitudes*, Bradford Books, MIT Press, 1983.

[Be] N.D. Belnap, A useful four-valued logic, in *Modern Uses of Multiple-Valued Logic* (eds. G. Epstein and J.M. Dunn), Reidel, 1977.

[Ca] J.A.K. Cave, Learning to agree, *Economics Letters* **12**, 1983, pp. 147-152.

[ChM] M. Chandy and J. Misra, How processes learn, *Proceedings of the 4th ACM Symposium on Principles of Distributed Computing*, 1985, pp. 204-214.

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

[Cr1] M.J. Cresswell, *Logics and Languages*, Methuen and Co., 1973.

[Cr2] M.J. Cresswell, *Structured Meanings*, MIT Press, 1985.

[DM] C. Dwork and Y. Moses, Knowledge and common knowledge in a Byzantine environment I: crash failures, to appear in *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[Eb] R. A. Eberle, A logic of believing, knowing and inferring, *Synthese* **26**, 1974, pp. 356-382.

[FH] R. Fagin and J.Y. Halpern, Belief, awareness, and limited reasoning, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985, pp. 491-501.

[FHV] R. Fagin, J.Y. Halpern, and M.Y. Vardi, A model-theoretic analysis of knowledge, *Proceedings of the 25th Annual IEEE Symposium on Foundations of Computer Science*, 1984, pp. 268-278.

[FV1] R. Fagin and M. Y. Vardi, An internal semantics for modal logic, *Proceedings of the 17th ACM Symposium on Theory of Computing*, 1985, pp. 305-315.

[FV2] R. Fagin and M.Y. Vardi, Knowledge and implicit knowledge in a distributed environment, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[FI] M.J. Fischer and N. Immerman, Foundations of knowledge for distributed systems, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[GP] J. Geanokoplos and H. Polemarchakis, We can't disagree forever, *Journal of Economic Theory* **28**:1, 1982, pp. 192-200.

[Ge] E. Gettier, Is justified true belief knowledge?, *Analysis* **23**, 1963, pp. 121-123

[GMR] S. Goldwasser, S. Micali, and C. Rackoff, The knowledge complexity of interactive proof-systems, *Proceedings of the 17th Symposium on Theory of Computing* 1985, pp. 291-304.

[Gr] J. Gray, Notes on data base operating systems, IBM Research Report RJ 2188, 1978.

[HF] J.Y. Halpern and R. Fagin, A formal model of knowledge, action, and communication in distributed systems: preliminary report, *Proceedings of the 4th ACM Symposium on the Principles of Distributed Computing*, 1985, pp. 224-236.

[HM1] J.Y. Halpern and Y.O. Moses, Knowledge and common knowledge in a distributed environment, *Proceedings of the 3rd ACM Conference on Principles of Distributed Computing*, 1984, pp. 50-61; revised version to appear as IBM Research Report, Jan. 1986.

[HM2] J.Y. Halpern and Y.O. Moses, A guide to the modal logics of knowledge and belief: preliminary report, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 480-490, 1985.

[HM3] J.Y. Halpern and Y.O. Moses, Towards a theory of knowledge and ignorance, in *Proceedings of the Workshop on Non-Monotonic Reasoning*, AAAI, 1984; also reprinted in *Logics and Models of Concurrent Systems* (ed. K. Apt), Springer-Verlag, 1985, pp. 459-476.

[HV] J.Y. Halpern and M.Y. Vardi, The complexity of reasoning about knowledge and time, unpublished manuscript, 1985.

[Hi1] J. Hintikka, *Knowledge and belief*, Cornell University Press, 1962.

[Hi2] J. Hintikka, Impossible possible worlds vindicated, *Journal of Philosophical Logic* 4 1975, pp. 475-484.

[HU] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, 1979.

[Ko] K. Konolige, Belief and incompleteness, SRI Artificial Intelligence Note 319, SRI International, Menlo Park, 1984.

[Kr] S. Kripke, Semantical analysis of modal logic, *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik* 9, 1963, pp. 67-96.

[Lad] R.E. Ladner, The computational complexity of provability in systems of modal propositional logic, *SIAM Journal on Computing* 6:3, 1977, pp. 467-480.

[LR] R. Ladner and J.H. Reif, The logic of distributed protocols, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[Lak] G. Lakemeyer, Steps towards a first-order logic of explicit and implicit belief, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[Leh] D.J. Lehmann, Knowledge, common knowledge, and related puzzles, *Proceedings of the Third Annual ACM Conference on Principles of Distributed Computing*, 1984, pp. 62-67.

[Len] W. Lenzen, Recent work in epistemic logic, *Acta Philosophica Fennica* 30, 1978, pp. 1-219.

[Lev1] H.J. Levesque, Foundations of a functional approach to knowledge representation, *Artificial Intelligence* 23, 1984, pp. 155-212.

[Lev2] H. J. Levesque, A logic of implicit and explicit belief, *Proceedings of the National Conference on Artificial Intelligence*, 1984, pp. 198-202; a revised and expanded version appears as FLAIR Technical Report #32, 1984.

[Lew] D. Lewis, *Convention, A Philosophical Study*, Harvard University Press, 1969.

[McH] J. McCarthy and P.J. Hayes, Some philosophical problems form the standpoint of artificial intelligence, in *Machine Intelligence* 4 (eds. B. Meltzer and D. Michie), Edinburgh University Press, 1969, pp. 463-502

[Me] M. J. Merritt, *Cryptographic Protocols*, Ph.D. Thesis, Georgia Institute of Technology, 1983.

[MZ] J.F. Mertens and S. Zamir, Formalization of Harsanyi's notion of "type" and "consistency" in games with incomplete information, C.O.R.E. Discussion Paper, Universite Catholique de Louvain, 1982.

[Mi] P. Milgrom, An axiomatic characterization of common knowledge, *Econometrica*, 49:1, 1981, pp. 219-222.

[MS] P. Milgrom and N. Stokey, Information, trade, and common knowledge, *Journal of Economic Theory* 26:1, 1982, pp. 17-26.

[Mon] R. Montague, Universal grammar, *Theoria* 36, 1970, pp. 373-398.

[Mo] R.C. Moore, Reasoning about knowledge and action, Technical Note 191, Artificial Intelligence Center, SRI International, 1980.

[MoH] R. C. Moore and G. Hendrix, Computational models of beliefs and the semantics of belief sentences, Technical Note 187, SRI International, Menlo Park, 1979.

[Mor] L. Morgenstern, A first-order theory of planning, knowledge and action, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[PR] R. Parikh and R. Ramanujam, Distributed processing and the logic of knowledge, *Proceedings of the Brooklyn College Workshop on Logics of Programs* (ed. R. Parikh), 1985, pp. 256-268.

[Pa] P.F. Patel-Schneider, A decidable first-order logic for knowledge representation, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985, pp. 455-458.

[PC] C.R. Perrault and P.R. Cohen, It's for your own good: a note on inaccurate reference, in *Elements of Discourse Understanding* (eds. A.K. Joshi, B.L. Webber, I.A. Sag), Cambridge University Press, 1981.

[Ra] V. Rantala, Impossible worlds semantics and logical omniscience, *Acta Philosophica Fennica* 35 1982, pp. 106-115.

[RB] N. Rescher and R. Brandom, *The Logic of Inconsistency*, Rowman and Littlefield, 1979.

[RSA] R. Rivest, A. Shamir, and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Communications of the ACM*, 21:2, 1978, pp. 120-126.

[Ro] S.J. Rosenschein, Formal theories of knowledge in AI and robotics, *Proceedings of Workshop on Intelligent Robots: Achievements and Issues*, SRI International, 1984, pp. 237-252.

[RK] S.J. Rosenchein and L.P. Kaelbling, The synthesis of digital machines with provable epistemic properties, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knoweldge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[Sa] M. Sato, A study of Kripke-style methods of some modal logics by Gentzen's sequential method, *Publications of the Research Institute for Mathematical Sciences, Kyoto University*, 13:2 1977.

[TW] T.C. Tan and S.R. Werlang, On Aumann's notion of common knowledge - an alternative approach, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[Th] R.H. Thomason, A model theory for propositional attitudes, *Linguistics and Philosophy* 4, 1980, pp. 47-70.

[Va1] M.Y. Vardi, A model-theoretic analysis of monotonic logic, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985, pp. 509-512.

[Va2] M. Y. Vardi, On epistemic logic and logical omniscience, *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge* (ed. J.Y. Halpern), Morgan Kaufmann, 1986.

[Wr] G.H. von Wright, *An Essay in Modal Logic*, North-Holland, Amsterdam, 1951.