RECENT ISSUES IN REASONING ABOUT KNOWLEDGE[1]

Rohit Parikh[2]

INTRODUCTION:

It is by now common knowledge that the recent period of intense activity in reasoning about knowledge begins with two books, both by philosophers: *Knowledge and Belief* by Hintikka [Hi], and *Convention* by David Lewis [Lew]. Since then there has been much talk about muddy children and non-coperating generals, but the field has actually grown to be quite wide and interesting and we will try and give an overview of some of the work that has been done and some of the issues that are still of concern.

It should be said at the outset that most of the work that has been done concerns itself with information rather than with knowledge. A familiar example from [FHV] goes "Dean doesn't know whether Nixon knows that Dean knows that Nixon knows about the Watergate break-in". Now it is likely in the case of Dean and Nixon that if they had the relevant information then they also had the knowledge, but of another, more recent president, it is easier to believe that he might have had the requisite information about the Iran-Contra affair and simply failed to make the necessary deduction which would lead to *knowledge*. This split between information and knowledge is one that almost all workers in the field have come to be aware of, but the issue is not resolved. More of this later.

It is customary to introduce a language for studying knowledge by starting with some base level "object" language, say propositional logic, and augmenting it with operators K_i where i is the name of some individual or knower and $K_i(A)$ stands for "i knows or has the information that A". One can then study the resulting language in terms of an abstract model theory and consider issues of satisfiability, validity and the computational complexity of the set of valid formulae.

The most common semantics used for this purpose is Kripke semantics where a set W of possible worlds is stipulated. There are some basic statements P_j, whose truth value is specified at each of these possible worlds, and some binary relations R_i, one for each knower i, are also specified. Roughly speaking, if s and t are two possible worlds (elements of W), then sR_it means that worlds s and t are indistinguishable to i, i.e. furnish the same evidence to i. Then the formula $K_i(A)$ holds at a world s iff A holds at each world t such that sR_it.

In the most common case, the relations R_i considered are equivalence relations, and if the base language is the propositional calculus, then the logic is S5-like, and the validity

[1]This paper should be regarded as a working draft towards a more comprehensive survey of recent work in reasoning about knowledge. If we have ignored important work, or misrepresented any position, we apologise in advance and welcome comments. When there are several references to the same result, we have tried to follow a *chronological* order.

[2]Department of Computer and Information Science, Brooklyn College of CUNY and Departments of Computer Science, Mathematics and Philosophy, CUNY Graduate Center, 33 W 42nd Street, New York, NY 10036. Email: RIPBC@CUNYVM.bitnet, Research supported in part by NSF grant CCR-8803409

problem is decidable, shown to be in DEXPTIME in [Pa] but is in fact in PSPACE. See [HV] for a survey of complexity results. A complete axiomatisation is also available, essentially along the lines of S5, see [Ma], [Pa].

While the Kripke style model theory is elegant and simple, it has two defects. One is that the notion of possible world as used there *includes* the psychological states of the knowers. Two worlds s and t may agree on all *facts*, but may be different because some knowers know different things in them. If this is the case, then the notion of possible world as in the Kripke models is *complex* and needs to be analysed. Another reason for dissatisfaction is that while the Kripke semantics yields a finite model property, it is in fact true, as shown by [FHV], that the state of total ignorance, where no one knows anything, can only be represented by an infinite Kripke model.

[FHV], as also [MZ],[3] therefore propose another model theory, where a possible world is represented as a tower of levels, the bottom level being facts, the next level being the individuals' knowledge of these facts, the next one specifying the individuals' knowledge of other individuals' knowledge, and so on. [FHV] prove that their version is equivalent to Kripke semantics and that the models are inter-translatable in a natural way.

The semantics that we have considered so far is abstract; an application, say to distributed computing, must give us some way of going from a described protocol to the relevant Kripke model, and the most common proposal, due independently to [PR], [CM] and [HM2], is to use the distinction between local runs and global runs as the basis for the relations R_i. Specifically, when processes co-operate in a computation, then each process has certain events happening locally as well as messages received from and sent to other processes. If A is a property of global states and process i has gone through a sequence r_i of local events (including sends and receives if any), then i *knows* A iff A is true in all global states compatible with r_i. [HV] point out that there are many choices that may be made in such models, whether time is synchronous or asynchronous, whether the processes have finite or infinite memory, etc., and they show that the complexity of the resulting logics can vary from PSPACE to Π^1_1-complete.

There have been some specific applications of knowledge in distributed computing. [PR] contains some preliminary results, but there are other, more substantial applications; e.g. in [HZ] to the sequence transmission problem, in [DM] to Byzantine failures and in [Maz] to the problem of recovery from crashes. See also [MT].

COMMON KNOWLEDGE:

Common knowledge as an issue in knowledge theory was introduced independently by Lewis [Lew] and Schiffer [Sch]. What they pointed out was that co-ordinated action, or proper communication, requires infinitely many levels of knowledge in the following sense: if, say, i and j are the two individuals involved, then there is in these situations some proposition

[3]The two models differ a little in that the [MZ] model includes probabilities, but we shall not go into this here.

A such that the infinitely many propositions $K_i(A)$, $K_j(A)$, $K_i(K_j(A))$,... must all be true for some co-ordinated action to take place, or respectively, some reference to be made successfully from i to j. [CM2] contains a series of amusing examples about a Marx Brothers movie to substantiate this claim.

The problem with common (or mutual) knowledge is that it seems difficult to attain. [HM] show that it cannot be attained in asynchronous systems and [CM2] give informal arguments why it cannot be attained in ordinary social situations, thus turning the whole business into something of a paradox.

Barwise [Bar] takes the stance that we really have *three* notions here: (i) the infinite iteration mentioned above, (ii) a fixed point B defined by $B = A \wedge K_i(B) \wedge K_j(B)$, and (iii) the existence of a situation s such that s implies A *and* that both i and j see s.[4] Certainly, many situations where we would attribute common knowledge *are* such as described in (iii). E.g. where both i and j are in the presence of a card which is lying face up on the table. However, common knowledge of *abstract* facts, needed to make use of the concrete facts, seems harder to explain this way.

Anyway, Barwise seems to be arguing that common knowledge may imply a transfinite iteration in certain cases, a point of view which receives some support from a result proved in [Pa4] that some facts can be learned only through transfinite dialogues and cannot be learned at any finite stage. However, a willingness to take a risk, no matter how small, reduces the needed time to a finite value. Perhaps this explains why we might, in practice, make do without actual common knowledge. A dance requires common knowledge between partners, but only approximate common knowledge is in fact present, and an occasional stubbed toe is the price that most of us are willing to pay for the pleasure.

Common Knowledge of a fact is in some sense the highest level at which a group might know it. The lowest level (barring implicit knowledge) is where the fact is known to *just one* of the individuals in question. [Pa2] considers the question of what intermediate levels there might be and shows[5] that they correspond precisely to a family of regular languages in the alphabet $\{K_1, ..., K_n\}$ where $\{1, ..., n\}$ are the individuals involved.

However, is there only *one* level of common knowledge? Suppose, for example, that a young man sitting next to a girl moves his knee so that it touches hers. This fact will then be common knowledge between them even if she ignores it. If, however, she says: "Excuse me, but your knee is touching mine", then the dynamics between them will change and whether he removes his knee or not, the action will have a different significance than if she had not spoken. A similar situation will arise if country A masses troups on the border of country B. If it is common knowledge that country B has a good espionage system, then this massing of troups will be common knowledge between A and B. The situation will nonetheless change if the prime minister of country B summons the ambassador of country

[4]Barwise suggests that the correct representation of cases (ii) and (iii) requires us to resort to non-well founded sets in the sense of Aczel [Ac].

[5]This result is joint with P. Krasucki.

A and mentions the massing of troups. The common knowledge will rise to a "higher" level and some action will now become necessary.

This seems to indicate that what we call common knowledge may actually stand for a game-theoretic situation, that such situations may differ from each other, and may neither imply nor require common knowledge as we *usually* understand it. This is already implicit in [Lew], but clearly there is a great deal of subtlety here.

STARTING FROM IGNORANCE:
Suppose I tell you that $0 < a < b$ and that $a \cdot b = 6$, then you know that $a = 2$ and $b = 3$. If, however, I had told you instead that $a \cdot b = 12$, then you would not know what a and b are. It turns out that your ignorance in the second case cannot be proved in a monotonic logic, since it will not survive the additional (and consistent) information that a is even. The point is that your ignorance is due to the implicit assumption that what you have been told about a and b is *all* you know about them. McCarthy has suggested that this kind of default reasoning be formalised using the inference rule

$$\frac{\Gamma \not\vdash K_i(A)}{\Gamma \vdash \neg K_i(A)}$$

This rule, however, does have its problems. For example, the formula $C = K_i(A) \vee K_i(B)$ does not imply $K_i(A)$ nor $K_i(B)$ and hence implies their negations by McCarthy's rule. These, however, together imply the negation of the original formula. A more subtle argument shows that even the empty set Γ is inconsistent under McCarthy's rule. This problem is tackled in [Pa] where a model theory and a completeness theorem for McCarthy's rule are given. Roughly, the idea is that larger Kripke models represent more possibilities and hence less knowledge. Hence a state of maximum ignorance, compatible with certain given facts, is represented by a largest Kripke model, and the existence of such a model is equivalent to consistency under certain normal applications of McCarthy's rule. When consistency does obtain, then all formulae true in the largest Kripke model can be proved through normal[6] applications of McCarthy's rule. In particular, the formula C above *is* inconsistent, it has no largest model, but the empty set (thank heavens!) turns out to be just fine.

Normal deducibility from consistent formulae can be shown to be in PSPACE, but, as Joe Halpern has pointed out, the existence of a largest Kripke model for a given formula A may be non-elementary.[7]

[6] An application of McCarthy's rule to $K_i(A)$ is *normal* if all subformlae of A, to which the rule could be applied, have been dealt with first.

[7] In particular, while the completeness theorem, as implied by Theorem 8(i) of [Pa], is correct, there are subtle errors in the proofs of parts (ii) and (iii), and all we can say is that the consistency problem is decidable and, at worst, of the same complexity as the system WS1S.

THE PROBLEM OF LOGICAL OMNISCIENCE:

One of the principal diffferences between knowledge and mere information shows up in the fact that if we have information that A and information that $A \rightarrow B$, then we also have information that B. Moreover, if B is logically true, then it requires no information in the first place. If an individual's knowledge *does* happen to have these closure properties, then we call that individual logically omniscient. Nonetheless, it does happen in fact that we are not logically omniscient and that we often fail to know B, either because the computation is intractable, or because we happen not to think of the justification for B, or, as Doyle [Doy] points out, we are not actually interested in B.

One area where this issue becomes important is public key cryptography, where the cypher-text does contain the same information that the plain-text does, but, lacking the key, the computation of the plain-text from the cypher-text is intractable. Thus a theory of knowledge which avoids assuming logical omniscience is crucial.

There have been several attempts to deal with this problem by developing logics in which knowledge is not closed under all logical inferences. Some examples include [FH], [Mos] and [FZ]. These papers all attempt to develop logics of knowledge which allow for limited reasoning powers. However, the analysis of why the systems proposed are the right ones is not completely convincing and the logics go only part way towards the heart of the logical omniscience problem. By contrast, the papers [Doy] and [Pa3] contain informal analyses of the problem and give us some insight into it, but there are no formal systems proposed that one could use to formalise actual, limited, knowledge. Indeed, [Doy] argues that formal systems are bound to be distortions, since they do not take into account the *goals* of the reasoning agent or the fact that resources, while bounded, may change with time.

It is worth mentioning the beautiful results in [Va] where Vardi analyses the logical omniscience problem from a pure complexity point of view and shows that the complexity of deducing logical conclusions stems from the ability to put together two distinct known facts. In other words, it is the *binary* rules of inference which are computationally expensive and, perhaps, account for why we don't know as much as we should.

A related, interesting problem is that of formalising the logical *goals* of a public key cryptographic system or of a zero knowledge proof system. What the standard literature gives us is the implementations of the goals, "I should convince you that I can factorise n without actually giving you the factors", but we do not have a formal language for stating the specifications. [BAN] contains a formal language that represents a beginning in this direction, but it would be nice to have a clean language whose formal semantics corresponds to our intuitions about what the logical issues are and which allows us to separate the complexity issues from the logical ones.

THE SYNTACTIC APPROACH:

In his "Three Grades of Modal Involvement", [Q], Quine proposed that modalities might

apply to sentences rather than to propositions. This can of course also be done with knowledge operators and then this device neatly bypasses the issue of referential opacity. For now it is easy to see why I might know the sentence $\ulcorner A \urcorner$ and fail to know the sentence $\ulcorner B \urcorner$, even though the propositions A and B are logically equivalent.

Unfortunately, this approach has its limitations. Thomason [Th] showed, using techniques adapted from Montague, that very reasonable systems following this approach and containing a certain amount of arithmetic, are inconsistent. There is more recent work in this direction, by Asher and Kamp [AK], using techniques of Herzberger and Gupta and an abstract version of the basic problem by Koons [Koo]. While these approaches do not lead to any definitive logics, this line is still a promising one.

THE PROBLEM OF IDENTITY:

When one uses an ATM to withdraw money, the screen sometimes contains reference to an entity identified as "I". One could ask here whether this "I" is the terminal, the central computer that drives it, or perhaps the bank itself. We don't ask, since we do usually get the money and that is all that matters, but it is worth remembering that in a distributed computing situation, the *individuals* i who do the knowing are stipulated by us. [HM] refer to knowledge jointly held by several individuals as implicit knowledge. Thus if I know A and you know $A \rightarrow B$, then together, we implicitly know B. But this distinction between implicit and explicit knowledge presupposes that we know who the individuals are.

To consider one example, while we usually think of a Turing machine as an individual which attempts to compute (say) in polynomial time, we could also think of it as a *system* consisting of infinitely many tape squares together with one head, which have implicit (or distributed) knowledge whether the given string x is in the stipulated language L, but communication is needed to concentrate this implicit knowledge at a single node so that it can be used. If this is the case, then we should not regard all forms of implicit knowledge as equal, but differentiate between them according to how much communication is necessary to make it explicit. In this context one could well regard Yao's theory [Y] as a study of implicit knowledge.

FINAL REMARKS:

There are many issues that we have not been able to touch on here. Auto-epistemic reasoning is one. Applications to Mathematical Economics is another. There is also a large body of strictly philosophical literature dealing with the nature of knowledge and beliefs and the general issue of propositional attitudes. Hopefully some of these will be addressed by some of the other tutorials.

REFERENCES:

[Ac] P. Aczel, *Non Well-founded Sets.* CSLI Lecture note 14, 1988.

[AK] N. Asher and H. Kamp, The Knowers Paradox and Representational Theories of Attitudes, in *TARK-I*, Ed. J. Halpern, Morgan Kaufmann 1986, pp. 131-148.

[BAN] M. Burrows, M. Abadi and R. Needham, Authentication: A Practical Study in Belief and Action, in *TARK-2*, Ed. M. Vardi, Morgan Kaufmann 1988, pp. 325-342.
[Bar] J. Barwise, Three Views of Common Knowledge, in *TARK-2*, Ed. M. Vardi, Morgan Kaufmann 1988, pp. 369-380.
[CM] M. Chandy and J. Misra, How Processes Learn, *Distributed Computing* **1**:1, 1986, pp. 40-52.
[CM2] H. H. Clark and C. R. Marshall, Definite Reference and Mutual Knowledge, in *Elements of Discourse Understanding*, Ed. Joshi, Webber and Sag, Cambridge U. Press, 1981.
[DM] C. Dwork and Y. Moses, Knowledge and Common Knowledge in a Byzantine Environment, *TARK-I*, Ed. J. Halpern, Morgan Kaufmann 1986, pp. 149-170.
[Doy] J. Doyle, Knowledge, Representation and Rational Self-Government, in *TARK-2*, Ed. M. Vardi, Morgan Kaufmann 1988, pp. 345-354.
[FHV] R. Fagin, J. Halpern and M. Vardi, A Model-Theoretic Analysis of Knowledge (research report), RJ 6461, IBM 1988.
[FZ] M. Fischer and L. Zuck, Relative Knowledge and Belief, Research report YALEU/DCS/TR-589, 1987.
[Hi] J. Hintikka, *Knowledge and Belief*, Cornell U. Press, 1962.
[HM] J. Halpern and Y. Moses, Knowledge and Commmon Knowledge in a Distributed Environment, *Proc. 3rd ACM Symposium on Distributed Computing* 1984 pp. 50-61
[HM2] J. Halpern and Y. Moses, A Guide to the Modal Logics of Knowledge and Belief, *Ninth IJCAI*, 1985, pp. 480-490.
[HV] J. Halpern and M. Vardi, The Complexity of Reasoning about Knowledge and Time, *JCSS* **38** (1989) pp. 195-237.
[HZ] J. Halpern and L. Zuck, A Little Knowledge goes a Long Way, *Proc. 6th PODC*, 1987, pp. 269-280.
[Koo] R. Koons, Doxastic Paradoxes without Self-Reference, in *TARK-2*, Ed. M. Vardi, Morgan Kaufmann 1988, pp. 29-42.
[Lew] D. Lewis, *Convention, a Philosophical Study*, Harvard U. Press, 1969.
[Ma] D. Makinson, On Some Completeness Theorems in Modal Logic, *Zeit. f. Math. Logik* **12**, 1966, pp. 379-384.
[Maz] M. Mazer, A Knowledge Theoretic Account of Recovery in Distributed Systems, *TARK-2*, Ed. M. Vardi, Morgan Kaufmann 1988, pp. 309-324.
[Mos] Y. Moses, Resource-Bounded Knowledge, in *TARK-2*, Ed. M. Vardi, Morgan Kaufmann 1988, pp. 261-276.
[MT] Y. Moses and M. Tuttle, Programming Simultaneous Actions using Common Knowledge, Research Report MIT/LCS/TR-369 (1987)
[MZ] J. Mertens and S. Zamir, Formulation of Bayesian Analysis in Games with Incomplete Information, *Int. J. of Game Theory* **14** (1985) pp. 1-29.
[Pa] R. Parikh, Logics of Knowledge, Games and Dynamic Logic, *FST-TCS* 1984, Springer LNCS 181, pp. 202-222.
[Pa2] R. Parikh, Levels of Knowledge in Distributed Computing, *IEEE LICS Symposium*,

1986, pp. 314-321.
[Pa3] R. Parikh, Knowledge and the Problem of Logical Omniscience, *ISMIS-87*, North Holland, pp. 432-439.
[Pa4] R. Parikh, Finite and Infinite Dialogues, to appear in the Proceedings of a Workshop on Logic and Computer Science, MSRI, November 1989.
[PR] R. Parikh and R. Ramanujam, Distributed Computing and the Logic of Knowledge, *Logics of Programs* 1985, Springer LNCS 193, 256-268.
[Q] W. V. Quine, Three Grades of Modal Involvement, in *the Ways of Paradox*, Harvard U. Press, 1975. (Originally published in 1953).
[Sa] M. Sato, A Study of Kripke-type Models for Some Modal Logics, Research report, Kyoto University, 1976.
[Sch] S. Schiffer, *Meaning*, Oxford U. Press, 1972.
[Th] R. Thomason, A Note on Syntactic Treatments of Modality, *Synthese* **44** (1980), pp. 391-395.
[Va] M. Vardi, On the Complexity of Epistemic Reasoning, *4th IEEE-LICS Symposium*, 1989, pp. 243-252.
[Y] A. Yao, Some Complexity Questions Related to Distributed Computing, *Proc. 11th ACM-STOC*, (1979), pp. 209-213.