MICHAEL BACHARACH AND PHILIPPE MONGIN

EPISTEMIC LOGIC AND THE FOUNDATIONS OF GAME THEORY

For two reasons the meeting of epistemic logic and game theory was no doubt inevitable. In the first place, during the 1970s and 1980s game theorists were developing, using their own formal tools, more and more precise treatments of epistemic matters. In the second place, it was becoming clear during these years that the formalisms of measure and probability in which game theory is standardly cast were much less distant than they at first seemed from modal propositional logic, the branch of logic to which the most widely used epistemic logics belong. In this foreword we shall briefly describe these twin developments, to which each of the five papers collected in this special issue of *Theory and Decision* attests in its own way. We shall take as read the main ideas of epistemic logic itself; readers who would like an introduction to the subject may wish to consult Section 2 of Bacharach's paper.

The many attempts which have been made either to justify the use of Nash equilibrium as a solution, or to restrict it by appropriate 'refinements', have enabled game theorists to detach, and explore the game-theoretical consequences of, alternative formal principles governing the subjective reasoning of the players.[1] To give one example, the analysis of Selten's 'subgame-perfect' equilibrium refinement has revealed unsuspected obstacles to classical arguments by 'backward induction'. Game theory has also moved in the opposite direction, seeking to weaken the Nash equilibrium solution concept in various criteria of 'rationalizability' which appeared in the 1980s. These criteria are justified by considerations which are distinctively epistemic. They incorporate a finite or infinite regress of reciprocal beliefs, rooted in simple beliefs in the rationality of the players and the rules of the game. Issues such as these, concerning the relation between solution-hood and the epistemic principles informing the reasoning of players, are central in the papers of Bacharach and Stalnaker.

Bacharach's paper offers an introduction to the basic ideas of modal propositional logic and standard epistemic logic as well as to certain

Theory and Decision **37**: 1–6, 1994.
© 1994 *Kluwer Academic Publishers. Printed in the Netherlands.*

relevant developments in 'nonmonotonic' logic.[2] He applies the concepts of epistemic logic to define a formal object called a *broad theory of a game*, intended to make manifest the full epistemic structure of the constructions used by game theorists to describe games. This formal object allows him to make explicit first the 'knowledge base' attributed to each player by the theorist, then the dual principle according to which players know all the logical consequences of the base ('cleverness') and only these consequences ('cloisteredness'). The first half of the principle is an assumption of 'logical omniscience' and raises the question of how we might restrict the deductive consequence relation under which standard epistemic logics assume belief sets to be closed. The second half, which is trickier to capture because it calls for a metalinguistic formulation, raises the opposite question of how we might expand the total collection of beliefs attributable to players; one way Bacharach proposes is the method, characteristic of theories of belief revision[3] and of nonmonotonic logics, which consists in allowing inferences to exceed, in appropriate cases, that which is sanctioned by classical deductions.

Stalnaker analyses alternative concepts of game-theoretic equilibrium by a new method which draws both on expected utility theory and on one of the classic constructions of modal logic, Kripke's (1963) semantics. Under the name *model of a game* he defines what is, in effect, a Kripke structure enriched by endowing each player with a prior probability measure and a decision function, each defined over the structure's set of possible worlds. Different versions of the notion of rationalizable solution, and Nash equilibrium, can be characterized, extensionally, by appropriate classes of models of the game. These classes correspond to natural epistemic properties of the players. For example, Stalnaker characterizes in this manner the class of Nash equilibria in two-person games thus: P*i* knows the beliefs of P*j* about P*i*'s strategy choice, and knows that P*j* maximizes expected utility. Stalnaker then uses his semantic method to define and to justify epistemically a new solution concept, 'strong rationalizability'.

Both Bacharach's and Stalnaker's papers concern the theory of games of 'complete information'. A now classic construction in the theory of games of 'incomplete information' illustrates the role played by players' beliefs in another way. In 1967–68 Harsanyi showed that an

adequate 'Bayesian' definition of such games required the explicit introduction of an infinite regress of reciprocal probabilistic beliefs. This regress is rooted in simple beliefs about the types of the players, that is, about the parameters which describe their utility functions rather than their rationality and the rules of the game, as in the theory of rationalizability. The present collection can not, unfortunately, do justice to the issues surrounding such hierarchies of beliefs: it bears on them only indirectly, in the analyses of common knowledge contained in the contributions.

The second development which has led to the *rapprochement* of epistemic logic and game theory concerns the technical notions of belief and 'common belief' employed in the two disciplines. Aumann's (1976) definitions of these notions (or more exactly of their counterparts for knowledge) are expressed in the language of events and partitions of a state space, a language which is to the game theorist what prose was to M. Jourdain. The structure of classes of subsets of the state set can be easily represented in terms of logical ideas. It is, for instance, now well known – indeed it is a sort of 'folk theorem' – that the partitional model of knowledge adopted by Aumann corresponds (in the technical sense of a soundness and completeness theorem) to the epistemic logic *S5*, one of the most exigent of epistemic logics. It has become evident that a precise relationship of this sort subsists between two representations of knowledge concepts (or belief concepts) in general, one expressed in set-theoretical and intuitive terms, the other in a formal language. The logician's approach is in one way more general than that of the game theorist, since it includes both these representations (formalized as, respectively, a semantics and a syntactical or deductive system) and displays their correspondence.

The paper of Modica and Rustichini exploits this duality to offer a critique of the partitional definition of one-person knowledge. The fragility of this definition becomes clear once it is reformulated syntactically, since *S5* includes the axiom, highly questionable from the epistemological standpoint, of 'negative introspection'. The authors define 'awareness' as the condition of either knowing, or knowing that one does not know. They suggest weakening the questionable axiom to a property of symmetry in awareness; they then

use this property to demonstrate a new decomposition of the *S5* axiom-system.

Lismont and Mongin too exploit the syntax-semantics duality. Their main purpose is the axiomatic analysis of the concepts of common belief and common knowledge.[4] After contrasting the logical approach with the informal axiomatizations of these notions in game theory and economic theory, Lismont and Mongin review a number of systems of modal propositional logic which might serve their purpose. At a minimum, these systems require the individual belief operators to obey the monotonicity rule RM of modal logic,[5] and a common-belief operator to satisfy three requirements: a 'fixed point' axiom, a 'rule of induction', and RM. The systems differ according to which additional axiomatic constraints they impose on the operators. The system implicit in Aumann's definition is perhaps the strongest of those that could be reasonably be proposed. It is in fact a special case of a system of intermediate strength proposed by Fagin, Halpern, Moses and Vardi,[6] which imposes on the operators Kripke's classic system *K*. Lismont and Mongin express their preference for the minimal variant described above, and demonstrate its soundness and completeness with respect to the semantics of neighbourhood structures. The latter are a more general (and arguably more natural) semantics for modal logic than Kripke's.

This symposium could not, unfortunately, give proper expression to all the major trends in epistemic logic. For example, the theories of belief revision and of nonmonotonic logic are only touched on, and specialists in Artificial Intelligence will perhaps deplore the absence of the important and promising development of probability logics.[7] The systematic character which the syntax-semantics duality gives to the modal logical approach applies, for the time being, only to the qualitative (and thus, from one point of view, most trivial) aspects of probability and related concepts:[8] the structure of algebras of events, the properties of sets of measure one, and so forth. The formal language may be enriched to take account of quantitative aspects too. But such an extension means the inevitable sacrifice of certain finiteness constraints which the logician standardly imposes on a syntax. It must be performed with some delicacy if the syntax is not to be diluted into a mere paraphrase of familiar semantic ideas.

If the collection can not span the whole field of epistemic logic, it is understandable that it can make no claim to deal systematically with the exciting research which is today forging links between game theory and the theories of computability and algorithmic complexity,[9] for the latter theories are adjacent to epistemic logic rather than part of it. The authors of this research offer it as one response to the concerns expressed years ago by Simon (1957). The notion of a Turing machine and, more dramatically, that of a finite automaton, provide tractable and plausible, if preliminary, models of 'bounded rationality'. Theory and Decision might perhaps, when the time is ripe, devote another symposium to the strategic applications of these important concepts.

By treating Turing machines in the language of, and using the techniques of, modal propositional logic, Shin and Williamson link the two fields of enquiry of the logic of knowledge and of computability. The idea of constructing a modal logic of provability goes back to Gödel himself, and gave rise to Boolos's (1989) work on the unprovability of consistency. Shin and Williamson identify in *S4* the system which formalizes the 'knowledge' of a Turing machine. This knowledge is assumed to consist of a recursively enumerable set of propositions, which is closed under deduction and under the provability operator; this set is generated from a base of items of knowledge which is part factual and part logico-mathematical. The latter component includes a formal theory of arithmetic to which Gödel's incompleteness theorems apply; making essential use of this assumption, Shin and Williamson show among other things that the knowledge of Turing machines cannot obey the negative introspection axiom.

For the two guest editors there remains the pleasant duty of thanking *Theory and Decision*, and especially its Editor-in-Chief, for entrusting them with the task of compiling the present collection, and for the sympathetic collaboration that has been extended to them during its preparation.

NOTES

[1]See Myerson's (1991) text for a survey of the variants of Nash equilibrium.
[2]See, e.g., Makinson (1994).

[3]See, e.g., Gärdenfors (1988).
[4]Epistemic logic offers only a cursory analysis of the distinction between belief and knowledge: in the standard approach it represents the two notions by a single operator, and distinguishes them merely by the presence or absence of the 'axiom of truth' (T).
[5]RM allows one to infer $KA \rightarrow KB$ from $A \rightarrow B$, and is essentially equivalent to the rule RE together with the axiom schema $K(A \wedge B) \rightarrow (KA \wedge KB)$.
[6]See, e.g., Halpern and Moses (1992).
[7]See Fagin, Halpern and Megiddo (1988) and references in the survey by Bacchus (1990).
[8]Among these related concepts are Shafer's (1976) belief functions, often used in Artificial Intelligence.
[9]See, e.g., Cutland (1980) and references in Shin and Williamson's paper.

REFERENCES

Aumann, R. J.: 1976, 'Agreeing to Disagree', *The Annals of Statistics* **4**, 1236–1239.
Bacchus, F.: 1990, *Representing and Reasoning with Probabilistic Knowledge*, M.I.T. Press, Cambridge, Mass.
Boolos, J.: 1989, *The Unprovability of Consistency*, Cambridge University Press, Cambridge.
Cutland, N. J.: 1980, *Computability: An Introduction to Recursive Function Theory*, Cambridge University Press.
Fagin, R., Halpern, J. Y., and Megiddo, N.: 1988, 'A Logic for Reasoning about Probabilities', Technical Report RJ 6190, I.B.M., San José, Ca.
Gärdenfors, P.: 1988, *Knowledge in Flux*, M.I.T. Press, Cambridge, Mass.
Halpern, J. Y. and Moses, Y.: 1992, 'A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief', *Artificial Intelligence* **54**, 319–379.
Harsanyi, J.C.: 1967–1968, 'Games with Incomplete Information Played by Bayesian Players: Parts I, II, III', *Management Science* **14**, 159–182, 320–334, 486–502.
Kripke, S.: 1963, 'Semantical Analysis of Modal Logic I', *Zeitschrift für Mathematische Logik* **9**, 67–93.
Makinson, D.: 1994, 'General Patterns in Nonmonotonic Reasoning', in D. Gabbay and C. Hogger (eds.), *Handbook of Logic for Artificial Intelligence and Logic Programming, H: Monotonic and Uncertain Reasoning*, Oxford University Press, Oxford.
Myerson, R.B.: 1991, *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, Mass.
Shafer, G.: 1976, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, N.J.
Simon, H.A.: 1957, *Models of Man, Social and Rational*, Wiley, New York.

M. Bacharach
University of Oxford,
Institute of Economics and Statistics,
St. Cross Building, Manor Road,
OX1 3UL Oxford, U.K.

P. Mongin
CORE,
34 Voie de Roman Pays,
1348 Louvain la Neuve,
Belgium