Foreword

Wiebe van der Hoek
Department of Computer Science
the University of Liverpool
UK wiebe@csc.liv.ac.uk

Spring 2005

1. Introduction

It was in 2002 that the idea arose that the time was right for a journal in the area of reasoning about Knowledge, Rationality and Action; a journal that would be a platform for those researchers that work on epistemic logic, belief revision, game and decision theory, rational agency, planning and theories of action. Although there are some prestigious conferences organised around these topics, it was felt that to have a journal in this area would have lots of added value.

What such a journal would typically be a platform for, would be the kind of problems addressed by researchers from Computer Science, Game Theory, Artificial Intelligence, Philosophy, Knowledge Representation, Logic and Agents. Problems that address artificial systems that have to gather information, reason about it and then make a sensible decision about what to do next. It is for this reason, that I am very happy that *Knowledge, Rationality and Action* (KRA) now exists as its own Section at Springer/Kluwer. For the clear and obvious links that the scope of KRA has with Philosophy, it was decided that KRA would be launched as a series within the journal *Synthese*.

This book collects the first two issues of KRA. Its index shows that these first two issues indeed address its 'core business': all the chapters refer explicitly to knowledge, for instance, and rationality is represented by the many contributions that address games, or reasoning with or about strategies. Actions are present in many chapters in this book: whether they are epistemic programs, or choices by a coalition of agents, or moves in a game, or votes by the members of a jury. All in all, there is an emphasis on *Information* and a notion of *Agency*.

What is furthermore striking, is that almost all chapters study these concepts in a *multi* agent perspective. In no paper in this book we have an isolated decision maker that only reasons about his own information and strategies, but always this is placed in a context of other agents, with some implicit or explicit assumptions about the *Interaction*.

Finally, this volume demonstrates that 'classical' approaches to Information and Agency co-exist very well with more modern trends that show how Knowledge, Rationality and Action can achieve a broad and refreshing interpretation; there is a chapter on 'classical AGM-like' belief revision, but also two on a modern approach in the area of Dynamic Epistemic Logic. There is a chapter on von Neumann games, but also two that defend Evolutionary Game Theory, a branch of game theory that attempts to loosen the 'classical assumptions' about 'hyper-rational players' in games. There are chapters solely on logical theories, but also one that suggests how we can bridge the gap between symbolic and connectionist approaches to cognition. Finally, there is a chapter on voting that relativises Condercet's 'classical' Jury's Theorem.

In the next section I will briefly give some more details about the themes of this book. While this can be conceived as a 'top-down' description of the contents, in Section 3 I will give a more 'bottom-up' picture of the individual chapters.

2. Themes in this Book

The first two chapters (Logics for Epistemic Programs and A Counterexample to Six Fundamental Principles of Belief Formation), as well as the fourth (A Characterization of Von Neumann Games) and the last two (A SAT -Based Approach to Unbounded Model Checking for Alternating-Time Temporal Epistemic Logic (Chapter 9) and Update Semantics for Security Protocols) explicitly deal with *information change*. It is also subject of study in part of Chapter 3, Comparing Semantics of Logics for Multi-Agent Systems. Whereas the first chapter, in order to develop a Dynamic Epistemic Logic, uses a powerful object language incorporating knowledge, common knowledge and belief, the second chapter formulates and analyses six simple postulates, in a meta- and semi-formal language. A main difference in the two approaches is that in Chapter 2, the theory has to explain how 'expectations' implicitly encoded in a belief set determine the next belief (when a conjunction has to be given up, say), whereas in the first chapter, this is not an issue: there, the effect of the learning is exhaustively described by an *epistemic program*.

Moreover, in the first chapter, the 'only' facts prune to change are *epistemic* facts (for instance, one agent might learn that a second agent now *does* know whether a certain fact holds), where they are 'objective' in the second. Chapter 10 uses a mix of both kinds of information change: there, it can be both objective and 'first order'. Chapter 3 and 9 analyse the dynamics of knowledge in the context of Alternating Transition Systems, where we have several agents that jointly determine the transition from one state to another. These two chapters do not have operators that explicitly refer to the *change* of knowledge: it is encoded in the epistemic and action relations in the model. This is similar to how knowledge is dealt with in extensive games, where it is represented in the information partition of the underlying game trees. Chapter 3 demonstrates that, indeed, there is a close relation between Alternating Transition Systems and Concurrent Game Structures, of which the latter are often conceived as a generalisation of game trees.

The chapters 9 and 10 take a genuine Computer Science perspective on some of the issues analysed in chapters 1 and 3. More in particular, both chapters address the problem of (automatic) *verification* of complex agent systems. The systems of study in chapter 9 are the Alternating-time Transitions Systems of chapter 3, more in particular those which incorporate an epistemic component. The problem addressed in this chapter is that of *model checking* such systems: given a description of a transition system model, and a property expressed in an appropriate logic, can we automatically check whether that property is true in that model? Chapter 10's aim is to develop verification methods for the epistemic program-type of action of Chapter 1, in the area of *security protocols*. Since the use of encryption keys in such protocols is to hide information in messages from specific agents, but make it available to others, Dynamic Epistemic Logic seems an appropriate tool here.

Games, or at least, game like structures, are the object of study in the chapters 3, 4 and 5 (AN EVOLUTIONARY GAME THEORETIC PERSPECTIVE ON LEARNING IN MULTI-AGENT SYSTEMS). The fist of the two address knowledge or belief in such games, whereas the emphasis on the fifth chapter is more on *lack* of it. Chapter 4 studies memory of past knowledge for players playing a game in extensive form. If a player knows φ now, is he guaranteed to always know that he ever knew φ? The chapter gives a necessary and sufficient condition for it, and shows that players having perfect recall is closely connected to the notion of a von Neumann game.

The other two chapters (i.e., 3 and 5) have in common that they try to combine and relate different formalisms. Chapter 3 compares and relates several *semantics* for game like logics, including Alternating-Time Temporal Logic (ATL) and Coalition Logic. Chapter 5 rather relates three *disciplines:* evolutionary game theory, reinforcement learning and multi-agent systems. The key trigger for this work is the insight that in many realistic multi-agent systems one has to weaken the 'classical' game theoretic assumptions about 'hyper-rational' agents, by players referred to as 'bounded rational' agents, who only have partial information about the environment and the payoff tables, and who have to learn an optimal 'policy' by trial and error.

The sixth chapter questions exactly the same assumptions that classical game theory makes on 'hyper-rational agents' as are debated in Chapter 5. Chapter 6 addresses the question how to explain which equilibria are chosen in signalling games, games that try to shed a light on language use and language organisation. The chapter proposes to replace current explanations for such selection, which rely on strong assumptions about rationality and common knowledge (thereof) by the players, i.e., the language users, by explanations that are based on insights from evolutionary game theory.

Whereas Chapter 3 relates, on a technical level, several semantics for games like logics, and Chapter 5 makes a case to combine three disciplines in order to study the dynamics of rationality in multi-agent systems, Chapter 7 (MONOTONIC INFERENCE AND NEURAL NETWORKS) *uses* a semantics in one research paradigm, i.e., non-monotonic logic as a symbolic reasoning mechanism, to bridge a gap to another paradigm, i.e. to connectionist networks in the sub-symbolic paradigm Doing so, the chapter is a step to wards bridging the gap between symbolic and sub-symbolic modes of computation, thus addressing a long standing issue in Philosophy of Mind.

Finally, Chapter 8 (EVOLUTION OF CONVERSATIONAL MEANING AND CONVERSATIONAL PRINCIPLES) addresses the issue of rational decision making in a group, or voting. It discusses a classical result that says that, in the scenario of majority voting, if every juror is competent, the reliability that the group decision is correct, converges to certainty, if the group size increases. Thus, this chapter also sits in the multi-agent context, but rather than accepting that a the result of a joint action is given by some transition function, this chapter discusses the rationality of a specific way to merge specific actions, i.e., those of voting. Moreover, again, it appears that knowledge is crucial here, because the chapter proposes, rather that to assume independence of the voters given the state of the world, we should conditionalize on the *latest evidence*.

3. Brief Description of the Chapters

In LOGICS FOR EPISTEMIC PROGRAMS, by Alexandru Baltag and Lawrence Moss (Chapter 1), the authors take a general formal approach to *changes of knowledge*, or, better, *changes of belief* in a multi-agent context. The goal of their paper is to show how several *epistemic actions* can be explained as specific *update operations* on 'standard' Kripke state-models that describe 'static' knowledge. Updates describe how we move from one state-model to another, and an epistemic action specifies, how such an update 'looks like', for every agent involved in it. An example of an epistemic action is that of a *public announcement*: if φ is publicly announced in a group of agents A, the information φ is truthfully announced to everybody in A, and that this is done is common knowledge among the members of A. However, in more private announcements, it may well be that agent b learns a new fact, whereby c is aware of this, without becoming to know the fact itself. The approach of the authors is unique in the sense that they also model epistemic actions by Kripke-like models, called *action models*.

In A Counterexample to Six Fundamental Principles of Belief Formation, Hans Rott reconsiders six principles that are generally well accepted in the areas of non-monotonic reasoning, belief revision and belief contraction, principles of *common sense reasoning*. They can all be formulated just by using conjunction and disjunction over new information, or information that has to be abandoned. Rott then pictures a reasoner who initially 'expects' that from a set of possible alternatives a, b, or c, none will be chosen. He then sketches three possible (but different) scenarios in which the reasoner learns in fact that $a \vee b$, $a \vee b \vee c$ and c *do* hold, after all. Depending on how each new information is in line or goes against the reasoners 'other expectations', he will infer different conclusions in each scenario. Interestingly, the six principles are then tied up with principles in the theory of rational choice, most prominently to the Principle of Independence of Irrelevant Choices: one's preferences among a set of alternatives should not change (*within* that set of alternatives), if new options present themselves. Rott argues that, in the setting of belief formation, the effect of this additional information should exactly be accounted for and explained. The chapter concludes negatively: logics that are closer to modelling 'common sense reasoning' seem to have a tendency to drift away from the nice, classical patterns that we usually ascribe to 'standard logics'.

Alternating Transition Systems are structures in which each joint action of a group of agents determines a transition between global states of the system. Such systems have inspirations from as diverse as game theory, computation models, epistemic and coalition models. Chapter 3 (Comparing Semantics of Logics for Multi-Agent Systems, by Valentin Goranko and Wojciech Jamroga) uses these structures to show how (semantically) several frameworks to reason about the abilities of agents are equivalent. One prominent framework in their analyses is that of *Alternating-time Temporal Logic (ATL)*, a logic intended to reason about what coalitions of agents can achieve, by choosing an appropriate strategy, and taking into account all possible strategies for the agents outside the coalition. The chapter first of all shows that the three different semantics that were proposed for ATL are equivalent. Moreover, the authors demonstrate that ATL subsumes (Extended) Coalition Logic. Last but not least, they show that adding an epistemic component to ATL (giving ATEL), can be completely modelled *within* ATL, the idea being, to model the *epistemic* indistinguishability relation for each agent as a *strategic* transition relation for his 'associated epistemic agents'.

In Chapter 4, (A CHARACTERIZATION OF VON NEUMANN GAMES IN TERMS OF MEMORY), Giacomo Bonanno analyses knowledge, and the memory of it, in the context of extensive games with incomplete information. In such a game, it is assumed that each player can be uncertain about the nodes in which he has to make a move. First of all, Bonanno extends this notion to an information completion, in which a player can have uncertainties about *all* nodes, not just the ones that are *his.* For such games, he defines a notion of *Memory of Past Knowledge (MPK)* in terms of the structure and the information partition in that game. If a game only allows for uncertainty for every player at his decision nodes, the the analogue of MPK is called *Memory of Past Decision Nodes (MPD).* Syntactically, MPK is related to perfect recall: it appears to be equivalent to saying that, at every node, a player remembers everything he has known before: if he knew φ in the past, then now he knows that he previously knew φ. This notion is then connected to that of von Neumann games, which, roughly, are games in which all players know the time. The main result of the chapter tells us that an extensive form game with incomplete information allows for an information completion satisfying MPK if, and only if, the game is von Neumann satisfying MPD. Note that from this, it follows that if we start with a game with Memory of Past Decision Nodes, this game can be extended to a game with Memory of Past Knowledge, if and only if it is a von Neumann game.

Chapter 5 (AN EVOLUTIONARY GAME THEORETIC PERSPECTIVE ON LEARNING IN MULTI-AGENT SYSTEMS) by Karl Tuyls, Ann Nowé, Tom Lenaerts and Bernard Manderick, is a survey paper that argues how three currently loosely connected disciplines can use and contribute to each other's development. They first of all take the stance that crucial assumptions in 'classical game theory' make their applicability to multi-agent systems and the real world rather limited: in the latter, we cannot always assume that participants have perfect knowledge of the environment, or even of the payoff tables. Rather than assuming that players are 'hyper-rational', who correctly anticipate the behaviour of all other players, they propose players that are 'boundedly rational', who are limited in their knowledge about the game and the environment, as well as in their computational resources. Moreover, such players *learn* to respond better by *trial and error*, which adds to the dynamics of the multi-player, or multi-agent system. Given these assumptions about the partially known dynamic environment, is seems natural to assume that learning and adaptiveness are skills that are important for the agents in that environment. The chapter argues how *reinforcement learning*, a theoretical framework that is already established in single-agent systems, has to solve several technical problems in order to be applicable to the multi-agent case. The problem is that in such richer systems, the reinforcement an agent receives, may depend on the actions taken by the other agents. This absence of 'Markovian behaviour' may make convergence properties of reinforcement learning, as they hold in the single agent case, disappear. In order to fully understand the dynamics of learning and the effects of exploration in multi-agent systems, they propose to use *evolutionary game theory* in such systems, which adds a solution concept to the classical equilibria, namely that of a strategy being evolutionary stable. A strategy has this property if it is robust against evolutionary pressure from any appearing mutant strategy. Apart from giving several examples illustrating the main concepts, they show how evolutionary game theory can be used as a foundation for modelling new reinforcement algorithms for multi-agent systems.

Evolution of Conversational Meaning and Conversational Principles by Robert van Rooy (Chapter 6) questions exactly the same assumptions that classical game theory makes on 'hyper-rational agents'. This chapter addresses the question how to explain which equilibria are chosen in signalling games, games that try to shed a light on language use and language organisation. The chapter proposes to replace current explanations for such selection, which rely on strong assumptions about rationality and common knowledge (thereof) by the players, i.e., the language users, by explanations that are based on insights from evolutionary game theory, especially that of an evolutionary stable strategy. Rather than obtaining Nash equilibriua in language games by relying on almost reciprocal assumptions about mutual (knowledge of) rationality, or using a psychological notion of *salience* to explain selection of a so-called conventional equilibrium, the chapter shows how that equilibrium will 'naturally' evolve in the context of evolutionary language games. It also uses evolutionary game theory to explain how conventions that enhance *efficient* communication are more likely to be adapted than those that do not. Finally, this chapter shows how costly signalling can account for honest communication.

The aim of Chapter 7 (Monotonic Inference and Neural Networks), by Reinhard Blutner, is mainly a methodological one, i.e., to show that model-theoretic semantics may be useful for analysing properties of connectionist networks. Doing so, the chapter is a step toward bridging the gap between symbolic and sub-symbolic modes of computation, thus addressing a long standing issue in philosophy of mind. The chapter demonstrates first of all that certain activities of connectionist networks can be seen as non-monotonic inferences. Secondly, it shows a correspondence between the coding of knowledge in Hopfield networks, and the representation of knowledge in Poole systems. To do so, the chapter makes the latter systems *weight-annotated*, assigning a weight to all possible hypotheses in a Poole system. Then, roughly, links in the network are mapped to bi-implications in the logical system. In sum, the chapter contributes to its goals by encouraging us to accept that the difference between symbolic and neural computation is one of perspective: we should view symbolism as a high-level description of properties of a class of neural networks.

Chapter eight (A Model of Jury Decisions Where All Jurors Have the Same Evidence by Franz Dietrich and Christian List) addresses the issue of rational decision making in a group, or voting. The setting is a simple one: the decision is a jury's decision about a binary variable (*guilty* or *not*) under the assumption that each juror is competent (predicts the right value of the variable with a probability greater than 0.5). Under this scenario, Condorcet's Jury Theorem predicts that the reliability of a jury's majority decision converges to 1 if the size of the jury increases unboundedly. This holds under the assumption that different jurors are independent conditional on the state of the world, requiring that for each individual juror, a new independent view on the world is available. The authors propose a framework in which the jurors are independent on *the evidence*, rather than the world. This evidence is called the latest common cause of evidence of the jurors votes. This framework seems to have a realistic underpinning: a jury typically decides on the basis of commonly presented evidence, not on independently obtained signals about the world–the latter often not even being allowed for use in the court room. The chapter's jury' theorem then shows that the probability of a correct majority decision is typically less than the corresponding probability in the Condorcet's model. It also predicts that, as the jury size increases, the probability of a correct majority decision converges to the probability that the evidence is not misleading.

Chapter 9 (A SAT -Based Approach to Unbounded Model Checking for Alternating-Time Temporal Epistemic Logic), by M. Kacprzak and W. Penczek address the problem of (automatic) *verification* of complex agent systems. The systems of study in chapter 9 are the Alternating-time Transitions Systems of chapter 3, more in particular those which incorporate an epistemic component. The problem addressed in this chapter is that of *model checking* such systems: can we, given a description of a transition system model, and a property expressed in an appropriate logic (ATEL, an epistemic extension of ATL), automatically check whether that property true in that model? To do so, this approach fixes one of the semantics for ATL given in Chapter 4, to apply a technique from unbounded model checking to it. Then, for a given model T and ATEL-property φ, a procedure is given to express them as Quantified Boolean Formulas, which, using fixed point definitions, in turn yield purely propositional formulas. A main theorem of the chapter then states that φ is true of T if and only if the obtained propositional formula is satisfiable. Hence, model checking ATEL is reduced to a SAT-based approach, an approach that has computational advantages over model checking using for instance Binary Decision Diagrams.

The last chapter, chapter 10, (Update Semantics for Security Protocols by Arjen Hommerson, John-Jules Meyer and Erik de Vink) addresses verification of security protocols. This becomes more and more important in an era where agents send more and more private, secret or sensitive messages over an insecure medium. Decryption keys are introduced to make specific messages only readable to specific agents, which makes the need to reason about higher order information (knowledge about knowledge) in a multi-agent protocol obvious. This chapter takes three kinds of updates, or messages (or, in the terminology of Chapter 1, ‘epistemic programs’): the public announcement of an object variable, the private learning of a variable and the private learning about the knowledge of other agents about variables. The chapter first of all give a Dynamic Kripke Semantics for these specific actions, not unlike the semantics using action models as proposed in Chapter 1, and then puts this semantics to work to model and reason about two specific security protocols, in which encrypted messages are sent and received. This chapter might well be a first step to apply the model checking techniques described in Chapter 9 to the dynamic logic framework of Chapter 1, in the area of security and authorisation protocols.