Report on *The Logic of Rational Agency* Dagstuhl Workshop, January 21 – 25

Wiebe van der Hoek and Michael Wooldridge

Department of Computer Science
University of Liverpool,
Liverpool L69 7ZF, United Kingdom.
{wiebe,m.j.wooldridge}@csc.liv.ac.uk

September 18, 2003

1 Description

The notion of a rational agent is one that has found currency in many disciplines, most notable economics, philosophy, cognitive science, biology, social sciences and, most recently, computer science and artificial intelligence. Crudely, a rational agent is an entity that is capable of acting on its environment, and which chooses to act in such a way as to further its own interests. There is much research activity in the formal foundations of such agents and multi-agent systems. Many mathematical approaches to developing theories of rational agency have been developed, including decision theory, game theory, and mathematical logic. In this seminar, we focussed on logical approaches to rational agency.

There are three aspects to the study of logical approaches to rational agency:

1. Philosophy
2. Logical foundations
3. Application

The first aspect is concerned with the primarily philosophical questions of what rational agency is and how we might go about characterising it. Within the artificial intelligence and AI communities, one approach in particular has come to dominate – the view of rational agents as practical reasoners, continually making decisions about what actions to perform in the furtherance of their intentions and desires. This view of rational agents is largely seen as going hand-in-hand with the view of agents as intentional systems – systems that

may best be characterised in terms of mentalistic notions such as belief and desires.

The logical foundations aspect of the study is concerned with the extent to which these aspects of agents (practical reasoning and mentalistic notions such as beliefs and intentions) can be captured within a logical framework of some kind. There are many well-documented difficulties with using classical (first-order) logic to express these aspects of agency, and so a key component of the logical aspect is finding and appropriate logical framework within which to express an agent's (different kinds of) beliefs, goals, plans, intentions, and how his actions can affect them over time. Although much has been done on modelling such a attitudes in isolation, it is still not clear how easy it is to combine several of them into one framework, let alone if one changes the perspective to multi-agent system. From a technical point of view, the logics of choice for expressing these aspects are extremely complex, combining temporal, modal, and dynamic aspects in a single framework. The theoretical and meta-logical properties of such logics (computational complexity, expressive power, completeness results, theorem proving techniques) are not well understood.

Finally, the application aspect is concerned with how we might apply logical theories of agency in the construction of automated agents. Logical theories of agency can be used as (1) a specification language, (2) a programming language, and (3) a verification language. Viewed as a specification language, a logic of rational agency can be used to specify the desirable properties of a system that is to be built. The development of formal methods for specifying the desirable properties of computer systems is a major ongoing area of research activity in computer science, and the view of computer systems as rational agents brings a new dimension to this study. Executable logics have also been a major research topic in computer science, with the programming language PROLOG being perhaps the best-known example of an executable logic framework. While the kinds of logics used in the development of agent theory are typically much more complex than those which underpin languages such as PROLOG, there is nevertheless some potential for developing executable fragments of agent logics. Finally, an interesting issue is the extent to which a computer system can formally be shown to embody some theory of agency. It is an as yet open question how we might go about attributing attitudes such as beliefs, desires, and the like to computer programs. Verifying that a system implements some theory of agency is thus a major research issue.

The structure of the seminar reflected the discussion above:

1. Philosophical foundations

   What is rational agency? What are the right primitives (beliefs, desires, etc) for modelling rational agents? How do these primitives relate to one-another?

2. Logical foundations

What are the alternatives (e.g., classical logics, modal logics, first-order meta-logics, dynamic/action logics, deontic logics, temporal logics, ...) for modeling of the primitive components of rational agency? What are appropriate semantic frameworks for these logics (Tarskian model theoretic semantics? Kripke semantics? computationally grounded Kripke semantics? other approaches?) What are the relative advantages of these different frameworks? How do we combine these primitives into a single logic? What are the theoretical properties (expressive power, completeness, decidability/undecidability, computational complexity, proof procedures) of these combined logics? How do we use these logics to capture macro (non-atomic) aspects of rational agency, such as decision making (games, distributed utilities,...), communication, perception, collective action?

3. Application

   How can we use agent logics in the specifiation of agent systems? How can we manipulate or otherwise refine these specifications to generate implementations? Can we directly execute these logics, and if so how? How do we verify that implemented systems satisfy some theory of agency (deductive approaches, model checking, ...)?

2 Program and Abstracts

2.1 Monday 21th: Philosophy and Social Sciences

1,2 Barbara Grosz *Commitment and Decision-making in Collaborative Activity*

3 John Bell, *Enlightened Utility Maximizers*

Reason, it seems, undermines trust; at least if rationality is equated with utility maximization as in game theory. The alternative is to appeal to moral or to idealistic political considerations. However considerations of this kind cut little ice with hard-nosed game theorists. This talk suggests that reason and trust are, after all, compatible if a more enlightened form of utility maximization is adopted. On this view, it is rational to trust someone if it is "good for business" to do so. This commercial view of rationality involves the notion of acceptable mutual risk in return for mutual benefit, in an evolving context of trust, and with a view to future cooperation. The argument is illustrated by providing a solution to a well-known problem of game theory, namely backward induction in games called centipedes.

4 Frank Dignum (work with David Kinny and Liz Sonenberg), *From Desires, Obligations and Norms to Goals*

Traditional models of agents based on Beliefs, Desires and Intentions make little or no distinction between desires and goals, and the process whereby

goals arise from desires is given scant attention. In this paper we argue that the distinction between desires and goals can be an important one, particularly in a Multi-Agent System context, where other sources of individual motivation such as *obligations and* norms may be present. This leads us to propose an extended BDI architecture in which obligations, norms and desires are distinguished from goals and explicitly represented. In this paper we consider suitable logical representations for and properties of these elements, and describe the basic method of operation of the architecture, focusing on how goal generation and goal maintenance may occur.

5 John H Woods, (work with D. Gabbay), *Normative Models of Rational Agency*

A good deal of contemporary cognitive science seeks to provide principled descriptions of various kinds and aspects of rational behaviour, especially in beings like us or AI simulacra of beings like us. For the most part, these investigators presuppose an unarticulated common sense appreciation of the rationality that such behaviour consists in. On those occasions when they undertake to bring the relevant norms to the surface and to give an account of that to which they owe their legitimacy, these investigators tend to favour one or other of three approaches to the normativity question. They are (1) the analyticity or truth-in-a-model approach; (2) the pluralism approach; and (3) the reflective equilibrium approach.

All three of these approaches to the normativity question are irrecoverably flawed, never mind that the first two have some substantial (and often tacit) provenance among logicians, and the third has enjoyed a flourishing philosophical career.

Against these views, we propose a strong version of what might be called normatively immanent descriptivism. We attempt to elucidate its virtues and to deal with what appears to be its most central vulnerability, embodied in the plain fact that actual human behaviour is not uncommonly irrational.

6 Gerhard Lakemeyer, *Towards Bridging the Gap between Logic and Real Robots*

High-level robot controllers for realistic domains often need to deal with processes which operate concurrently, change the world continuously, and where the execution of actions is event-driven as in "charge the batteries as soon as the voltage level is low." Moreover, the controllers need to take into account that the actions of real robots typically have uncertain effects and that their sensors are noisy. While non-logic-based robot control languages address these issues in one way or another, they do not support projection, that is, the ability to reason about how the world evolves when actions are executed. Such reasoning is useful both during the design of

robot controllers and for the robot itself to guide its actions. Logic-based control language like Congolog were specifically designed to support reasoning about action, yet they lack most of the features mentioned above. In this talk, I will show how Congolog can be extended along these lines in order to bridge the gap between robot control languages which are based on logic and those which are not.

7 Mehdi Dastani, (work with J. Broersen, Z. Huang, J. Hulstijn, L. van der Torre *Conflict Resolution for Goal Generation: An Alternative Classification of Agent Types*
In this paper we discuss a generic component of a cognitive agent architecture that merges beliefs, obligations, intentions and desires into goals. The output of different components may conflict and the way the conflicts are resolved determines the type of the agent. For component based cognitive agents, we introduce an alternative classification of agent types based on the order of output generation among components. This ordering determines the type of agents. Given four components, there are 24 distinct total orders and 144 distinct partial orders of output generation. These orders of output generation provide the space of possible types for the suggested component based cognitive agents. Some of these agent types correspond to well-known agent types such as realistic, social, and selfish, but most of them are new characterizing specific types of cognitive agents.

2.2 Tuesday 22nd: Agents and Games

1,2 Johan van Benthem, *Logic and Games– A Tutorial*
Interfaces between logic and game theory come in the guise of 'logic games' in one direction, and 'game logics' in another. We present the latter, looking at games first as structures for modal logic, dynamic logic, and MU-calculus, reflecting the fixed-point character of game- theoretic equilibria. Next we briefly consider preference structure an dits associated logics. Then, we consider the more detailed information dynamics of stepwise moves through imperfect information games, showing how concrete game actions involve dynamic-epistemic logic and model changes.

3 Paul Harrenstein, *Relating Logical Notions to Game Theoretical Solutions*
Logical notions of consequence have frequently been related to game-theoretical solution concepts. The relations betweeen classical logical consequence and a player having a winning strategy in a two-player zero-sum game is particularly well-known. Here, I propose a concept of consequence based on the Nash-equilibrium solution concept.

Our research concerns a relation between a sequence of theories $(\Gamma_i)_{i\in\pi}$ and formulae ϕ in a propositional language, $(\Gamma_i)_{i\in\pi} \models_\pi \phi$. Here, π is a partitioning of the propositional variables of the language. Intuitively,

each block $i \in \pi$ represents a player that has control over the semantical values of the propositional variables occurring in i. The extensions of the formulae in each theory Γ_i are partially ordered by the inclusion relation and as such determine for each $i \in \pi$ an partial preorder over the valuations for the language. Viewed as preference orders, this means that each $i \in \pi$ aims at satisfying as strong a subtheory of Γ_i as possible by assigning semantical values to the variables in i. With prefernce orders and the powers of the players being specified, we are in a position to define the Nash-equilibria with respect to π and $(\Gamma_i)_{i\in\pi}$ as a, possibly empty, subset of the valuations. We define $(\Gamma_i)_{i\in\pi} \models_\pi \phi$ iff ϕ holds in all Nash-equilibria and investigate its formal logical properties.

4 Marc Pauly, *Logic for Social Software - In Praise of Ignorance and Against Individualism*

Coalitional power in multistage processes is modeled using effectivity frames, which link an effectivity function to every possible state of the world. Effectivity frames are general enough to capture, e.g., what groups of agents can bring about in extensive games of perfect and almost perfect information. Coalition Logic is used to describe effectivity frames, and the question of generating an extensive game satisfying a given specification is formulated as a satisfiability problem in Coalition Logic.

Using this logical reformulation, we show that the complexity of this implementation problem depends on two parameters: For individual specifications (i.e., specifications which only refer to the powers of individual agents), the problem is shown to be PSPACE-complete for single-agent systems and NP-complete for multi-agent systems. Furthermore, for multi-agent systems, generating an implementation with perfect information is PSPACE-complete, whereas generating an implementation with almost perfect information is NP-complete.

5 Boudewijn de Bruin *Backward Induction* In a centipede game, the game theoretic solution concept of "backwards induction" (BI) has all players go down at all their decision nodes. This strategy, however, is generally considered to be counterintuitive, or irrational, as one expects the players to try to go across as long as possible. Voilà the paradox!

I will investigate to what extent an epistemic analysis of BI can help dissolving the paradox. Although opinions diverge as to the exact epistemic characterization of the solution concept (some claim common knowledge of rationality (CKrat) is needed, others claim weaker conditions), many accept that such an analysis of the players' beliefs will help to rid us of the paradox by showing that the necessary epistemic conditions are highly unrealistic. The hidden assumption is that from the hierarchy of knowledge statements that CKrat consists of (K1rat2, K1K2rat1, etc.) a player

may "derive" beliefs about the strategy his opponent will go for (in casu, BI).

I will argue for the following two claims. (i) CKrat is not unrealistic at all. (ii) CKrat is not sufficient to determine (uniquely) a player's beliefs.

Ad (i). Playing a game (in some sense) commits you to the intention to win the game. This is common knowledge, and, I will show, entails CKrat.

Ad (ii). I will reason by means of a thought experiment of "public announcement" of rationality (i.e., before the game starts both players say: "I am a rational player," thus making it the rationality commonly known). I will show that the players are not necessarily forced to backwards induction; depending on the way they process information, they can go different ways in distilling beliefs.

6 Mike Wooldridge, (work with Wiebe van der Hoek) *Time, Knowledge, and Cooperation: Alternating-time Temporal Epistemic Logic and its Applications*

Branching-time temporal logics have proved to be an extraordinarily successful tool in the formal specification and verification of distributed systems. Much of this recent success stems from the tractability of the model checking problem for the branching time logic CTL. Several successful verification tools (of which SMV is the best known) have been implemented that allow designers to verify that systems satisfy requirements expressed in CTL. Recently, CTL was generalised by Alur, Henzinger, and Kupferman in a logic known as "Alternating-time Temporal Logic" (ATL). The key insight in ATL is that the path quantifiers of CTL could be replaced by "cooperation modalities", of the form $<< G >>$, where G is a set of agents. The intended interpretation of an ATL formula $<< G >> \phi$ is that the agents G can cooperate to ensure that ϕ holds (equivalently, that G have a winning strategy for ϕ). It turns out that the resulting logic very naturally generalises and extends CTL. In this talk, I will discuss extensions to ATL with *knowledge modalities*, of the kind made popular by the work of Fagin, Halpern, Moses, and Vardi. Combining these knowledge modalities with ATL, it becomes possible to express such properties as "group G can cooperate to bring about ϕ iff it is common knowledge in G that ψ". The resulting logic — Alternating-time Temporal Epistemic Logic (ATEL) — has a range of applications, which will be discussed in the talk. In addition, I will relate some preliminary experiments with ATEL model checking, which shares the tractability property of its ancestor CTL.

2.3 Wednesday 23rd: Logics and Proof Methods

1 Clare Dixon, (joint work with Michael Fisher, Wiebe van der Hoek, Ullrich

Hustadt, John-Jules Meyer, and Renate Schmidt) *Resolution-based Proof in Combined Modal and Temporal Logics*
KARO (Knowledge, Actions, Results, Opportunities) is a complex agent theory, involving propositional dynamic logic, S5 and KD modal logics, and several non-standard operators. Currently there are no proof methods for KARO, which are important for developing verification techniques. Our aim is to develop automated proof methods for the KARO framework by using existing calculi for combined modal and temporal logics, in particular for the fusion of the branching-time temporal logic CTL and multi-modal S5. First a core of the full KARO framework is identified and a satisfiability preserving translating into CTL fused with multi-modal S5 is given. Next we provide a resolution based proof system in this logic. Finally we consider extensions to the core using this approach.

2 Balder ten Cate *The Hybrid Logic of Epistemic Actions*

3 Ullrich Hustadt *Proof methods for the KARO framework*
We give a short overview of a method for realising automated reasoning about agent-based systems. The framework for modelling intelligent agent behaviour that we focus on is a core of KARO logic, an expressive combination of various modal logics including propositional dynamic logic, a modal logic of knowledge, a modal logic of wishes, and additional non-standard operators. The method we present is based on a translation of core KARO logic to first-order logic combined with first-order resolution. We discuss the advantages and shortcomings of the approach and suggest ways to extend the method to cover more of the KARO framework.

4 John Jules Meyer (work with Jan Willem Roorda and Wiebe van der Hoek) *Iterated Belief Change in Multi-Agent Systems*
We present a model for iterated belief change in a multi-agent system. We use a combination of modal (epistemic) and dynamic logic. Two core notions are: the expansion of the beliefs of an agent and the processing of new information by an agent. A history of the information received is maintained, tagged with a source it stems from. Essentially, when an agent receives information it adds it to the history, on the basis of a selection function. (It may also forget information from the history.) Depending on the specific selection function (which may be time-based, trust-based, or otherwise) the agent's (iterated) belief change has different properties. We claim that in this way of modelling we can accommodate the various approaches to (iterated) belief change found in the literature. As the selection function depends on the particular agent we allow for multi-agent systems with heterogeneous agents as to their belief expansion attitudes.

5 Leon van der Torre (joint work with M. Dastani) *A Qualitative Decision Theory with a Goal-based Representation Theorem*

In this talk we introduce a qualitative decision theory based on belief (B) and desire (D) rules. Goals are not first class citizens but defined indirectly. We show that every agent which makes optimal decisions – which we call a BD rational agent – acts *as if* it is maximizing its achieved goals. This goal-based representation theorem implies that an agent can be formalized or verified as a goal based reasoner even when it does not reason with goals at all. Qualitative decision theory does for goals what for example Savage's classical decision theory does for utility functions, and it can in this sense play a role in the formal foundations of Dennett's intentional stance.

6 Maria Fasli, *Interrelations between the BDI Primitives*
The study of formal theories of agents has intensified over the last decade since such formalisms can be viewed as providing the specifications for building agent-based systems. One such theory views agents as having beliefs, desires and intentions. The BDI paradigm provides us with the means of describing different types of agents; a desirable quality, since agent-based systems are employed in various domains with diverse characteristics and therefore different requirements. This is accomplished by adopting a set of constraints that describe how the three attitudes are related to each other, namely a notion of realism. Although, three such notions have been explored in the literature: strong realism, realism and weak realism, no systematic attempt has been undertaken to study other available options. In this paper we explore the dynamics and possible interrelations between the three attitudes and we propose notions of realism for heterogeneous BDI agents. We explore a more wide range of possibilities by considering the types of relations between accessible worlds. Moreover, we distinguish between two broad categories of agents according to the relation between beliefs and intentions: bold and circumspect. We explore several interesting notions of realism for such agents and we argue that these come close to the desiderata for rational BDI agents.

7 Ron van der Meyden **TTBA**

2.4 Thirsday 24th: From Specification to Implementation

1,2 Michael Fisher, *Executing Specifications of Rational Agents*

3 Alessio Lomuscio, (joint work with M Sergot) *On Specifications and Agents*
The design of complex multi-agent systems is increasingly having to confront the possibility that agents may not behave as they are supposed to. In addition to analysing the properties that hold if protocols are followed correctly, it is also necessary to predict, test, and verify the properties that would hold if these protocols were to be violated.

We investigate an extension of interpreted systems to model correct functioning behaviour of agents and of the system as a whole. We combine this notion with the standard epistemic notions defined on interpreted systems to provide a formalism to reason about knowledge that agents are permitted to hold under ideal functioning circumstances. We extend this by introducing a doubly-indexed operator representing knowledge that an agent would have if it were operating under the assumption of a group of agents functioning as intended. We investigate the completeness problem for the first formalism and discuss the issue for the more general one.

Finally, we illustrate how the formal machinery of deontic interpreted systems can be applied to the analysis of such problems by considering three variations of the bit transmission problem, a widely used scenario in knowledge representation.

4 Massimo Benerecetti, *Model Checking for Multi-Agent Systems*

5 Yves Lesperance *On the Semantics of Deliberation in IndiGolog - From Theory to Implementation*

6 Julian Bradfield, *Logical Development of Multi-Agent Computation Patterns*

7 Wiebe van der Hoek (work with Frank de Boer, Koen Hindriks and John-Jules Meyer) *Goal: Goal-Oriented Agent Language.*
A long and lasting problem in agent research has been to close the gap between agent logics and agent programming frameworks. The main reason for this problem of establishing a link between agent logics and agent programming frameworks is identified and explained by the fact that agent programming frameworks have not incorporated the concept of a *declarative goal*. Instead, such frameworks have focused mainly on plans or *goals-to-do* instead of the end goals to be realised which are also called *goals-to-be*. In this paper, a new programming language called GOAL is introduced which incorporates such declarative goals. The notion of a *commitment strategy* - one of the main theoretical insights due to agent logics, which explains the relation between beliefs and goals - is used to construct a computational semantics for GOAL. Finally, a proof theory for proving properties of GOAL agents is introduced. An example program is proven correct by using this programming logic.

2.5 Friday 25th: Software Development

1,2 Fausto Giunchiglia *TROPOS: A Methodology for Agent-oriented Software Development*

3 V.S. Subrahmanian, *Scaling Agents*

4 Emil Weydert, *Cognitive dynamics - the ranking perspective*

The modeling of belief states and belief change are major issues for agent theories. However, most existing approaches - based on belief sets and conditionalization - do not adequately reflect the real world complexity. In particular, for decision-taking and revision, we need epistemic preferences and reasonable strategies to change them. Qualitative approaches, like belief orderings, are simple - but coarse-grained and unable to support independency considerations. Quantitative approaches, like probability measures, are powerful - but cumbersome and unable to support full belief. Ranking measures, quasi-probabilistic surprise valuations in the tradition of Spohn, are an interesting alternative. Unfortunately, the classical strategy for revising rankings based on the minimal information paradigm is order-dependent. But we can solve this problem by a suitable, canonical parallel revision strategy (for sets). This constitutes a small but important step to a general theory of cognitive dynamics.

3 Evaluation

We think the workshop was very successful. We know that some collaborations have been intiated during the event. Moreover, the following two special issues came out as spin off from the workshop, both refering explicitly in a forword to the event at Dagstuhl:

W. van der Hoek and M.J.W. Wooldridge (eds),
Towards a Logic of Rational Agency, special issue of *Logic Journal of the IGPL*, **11**:2, 2003. see `http://www3.oup.co.uk/igpl/Volume_11/Issue_02/`

W. van der Hoek and M.J.W. Wooldridge (eds),
The Dynamics of Knowledge, special issue of *Studia Logica*, **75**:1, 2003.