LOGICAL PRINCIPLES OF RATIONAL AGENTS

M Fisher

Rapporteur: Dr B N Rossiter

LOGICAL PRINCIPLES OF RATIONAL AGENTS

Michael Fisher

Centre for Agent Research and Development
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, U.K.

EMAIL: M.Fisher@doc.mmu.ac.uk
PERSONAL: http://www.doc.mmu.ac.uk/STAFF/M.Fisher
GROUP: http://www.card.mmu.ac.uk

September 2000

Contents

Lecture One

- Characterising rational agents
  - — selected characteristics of (rational) agents
  - — using rational agents
- Logical foundations
  - — modal logics
  - — temporal logics
- Logical theories of rational agents
  - — combinations of logics
  - — examples and agent theories

MICHAEL FISHER

RATIONAL AGENTS

What is an Agent?

Because of the popularity of agent-based systems, the question "what is an agent" has stimulated much discussion across a range of areas.

I will just consider the concept of an 'agent' as being a useful abstraction for software components that occur within complex, dynamic systems.

While the key element of an object is *encapsulation*, the key additional element of an agent can be seen as *autonomy*. Thus,

> an agent has the ability to act independently from, and irrespective of, its environment.

Hence:

you can't tell an agent what to do – you can only ask it!

Rational Agents (1)

However, we want more than just basic autonomous agents — we require agents that are *flexible*.

Why? Because:

- environments are dynamic and unpredictable
- agents are under varying real-time constraints
- agents may learn/evolve new behaviour
- agents are part of an open system (i.e. no fixed topology) and are under no central control

Thus, the type of agents we are interested in are capable of *flexible autonomous action* — these are termed *rational* (or *intelligent*) agents.

Such rational agents must be able to adapt their autonomous behaviour to cater for the dynamic aspects of their environment, their requirements and their knowledge.

Rational Agents (2)

In order to provide such flexibility, the elements, in addition to autonomy, that we typically require from rational agents are

1. pro-activeness
   i.e. the agent is not driven solely by events and so it takes the initiative and generates, and attempts to achieve, its own goals
2. social ability
   i.e. the agent interacts with other (sometimes human) agents and cooperates with these in order to achieve some of its goals
3. deliberation
   i.e. the agent can reason about its current state and can modify its subsequent actions and future goals according to the situation

Representing Rational Agents

Many formal models of rational agency share similar elements, in particular

- a *dynamic* element, allowing the representation of the agent's basic dynamic activity,
- an *informational* element, representing the agent's database of information,
- a *motivational* element, often representing the agent's goals, and
- a mechanism for *deliberation* that characterises the way in which motivations develop dynamically.

Such elements provide the characteristics we are interested in. For example:

- internal motivations for taking particular choices are required in order to provide pro-activeness;
- information about the agent's environment, its capabilities and other agents are all useful for social interaction in dynamic settings;
- deliberation is necessary in order to decide which motivations to adopt, and which items of information to (re-)consider.

Example: Spacecraft Landing

Imagine an agent controlling a spacecraft that is attempting to land on a planet.

The agent has

- information about the terrain of the planet
- information concerning target landing sites
- motivations, such as
  - to land soon
  - to avoid mountains
  - to remain aloft until safe to land
  - etc...

The agent must dynamically

- assess its information for veracity and, if necessary, revise the information held
- deliberate over (possibly conflicting) goals in order to decide what actions (for example, movement) to take, and
- based on its current state, generate new goals (for example to land near a particular target site) or revise its current goals

Using Rational Agents

While most of the agents developed, for example for the INTERNET, remain relatively simple, there are beginning to be applications where rational agent technology is used.

Examples include traffic control, resource management, business process modelling, real-time process control and telecommunications.

High profile examples include

- real-time fault monitoring on space shuttle
- air traffic control at Sydney airport
- real-time fault monitoring and diagnosis carried out in the Deep Space 1 mission

Towards Formal Logical Representation

The elements of rational agent theories are typically represented logically as follows

Dynamism — temporal or dynamic logics;

Information — modal logics of belief or knowledge;

Motivation — modal logics of goals, intentions or desires.

While it may seem peculiar to characterise software components in terms of mental notions such as belief and desire, this follows a well known approach termed the *intentional stance*.

Attributing such mental notions to agents provides us with a convenient and familiar way of describing, explaining, and predicting the behaviour of these systems.

Thus, the intentional stance simply represents an abstraction mechanism for representing agent behaviour.

Next we will examine the logical background for such representations.

References

Definitions and Theories:

Dennett — "The Intentional Stance". MIT Press, 1987

Cohen, Levesque — "Intention is Choice with Commitment". *Artificial Intelligence* 42, 1990.

Wooldridge, Jennings — "Intelligent Agents: Theory and Practice". *Knowledge Engineering Review* 10(2), 1995.

Applications:

Rao, Georgeff — BDI Agents: from theory to practice". *Proc. ICMAS*, 1995.

Jennings, Wooldridge (editors) — "Agent Technology: Foundations, Applications, and Markets". Springer-Verlag, 1998.

Muscettola, Pandurang Nayak, Pell, Williams — "Remote agent: To boldly go where no AI system has gone before". *Artificial Intelligence* 103(1-2), 1998.

LOGICAL FOUNDATIONS

Modal Logics

In classical logic, formulae are evaluated within a single fixed world.

For example, a proposition such as "it is Monday" is either *true* or *false*.

Propositions are then combined using constructs such as 'and', 'if....then', 'or', 'not', the constant symbols 'true' and 'false', and a set of symbols representing atomic propositions.

In modal logics, evaluation occurs within a range of *possible* worlds.

MICHAEL FISHER

Syntax and Semantics

As well as syntactic operations for manipulating the truth and falsehood of propositions within a world, operators for navigating between worlds are required.

In standard modal logics there are two such operators:

- '$\Box\varphi$' means that φ is true in all worlds accessible from the current world.
- '$\Diamond\varphi$' means that φ is true in *some* world accessible from the current world.

Thus, the truth of propositions is dependent upon the world in which they are evaluated.

But what does 'accessibility' mean? Its meaning is dependent upon the context in which the logic is to be used.

For example, all of the following interpretations for $\Box/\Diamond$ are common.

- is necessary/is possible
- believes/doesn't believe opposite
- knows/doesn't know opposite
- always in future/sometime in future

Semantics

Formulae are interpreted within models ($\mathcal{M}$) comprising

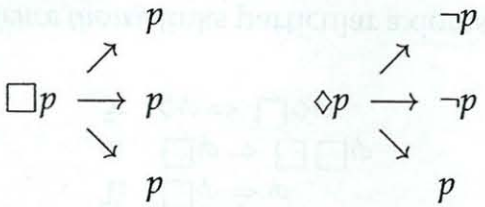
- a set of worlds, W,
- a binary relation, R, on worlds in W, and,
- a propositional interpretation π of type

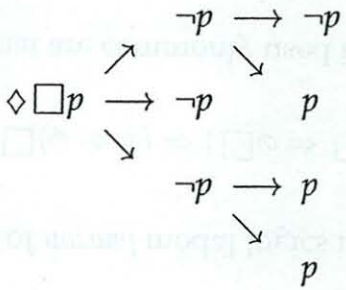
$$\pi: W \times \text{PROP} \rightarrow \{T, F\}$$

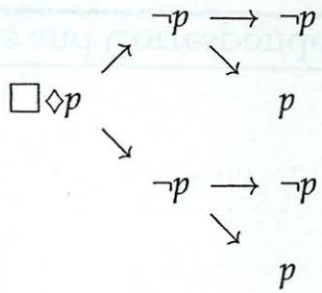
Thus, key part of the semantics is

$\mathcal{M}, w_1 \models p$	iff	$\pi(w_1, p) = T$
$\mathcal{M}, w_1 \models \Box\varphi$	iff	for all w_2, if $R(w_1, w_2)$, then $\mathcal{M}, w_2 \models \varphi$
$\mathcal{M}, w_1 \models \Diamond\varphi$	iff	there exists a w_2, such that $R(w_1, w_2)$ and $\mathcal{M}, w_2 \models \varphi$

Examples







Constraining Accessibility Relations

In modal logics, the properties of the accessibility relation, R, play a crucial role.

So far we have considered unrestricted relations.

If we now restrict the relation, we can induce interesting (and useful) effects in the logic used.

There are *many* properties of R, for example

- **reflexivity**: if $w_1 \in W$ then $R(w_1, w_1)$
- **transitivity**: if $R(w_1, w_2)$ and $R(w_2, w_3)$ then $R(w_1, w_3)$

Axioms and Correspondences

The core axiom of *normal* modal logics is the 'K' axiom:

K: $\Box(\varphi \Rightarrow \psi) \Rightarrow (\Box\varphi \Rightarrow \Box\psi)$

Other axioms that are commonly used in modal logics include

D: $\Box\varphi \Rightarrow \Diamond\varphi$
T: $\Box\varphi \Rightarrow \varphi$
4: $\Box\varphi \Rightarrow \Box\Box\varphi$
5: $\Diamond\varphi \Rightarrow \Box\Diamond\varphi$

Correspondence theory links particular axioms to properties of R, for example

- the 'T' axiom corresponds to reflexivity of R,
- the '4' axiom corresponds to transitivity of R.

Useful Axiom Combinations

The logic comprising axioms K, D, 4 and 5 (unsurprisingly called 'KD45') is commonly used where accessibility represents 'belief'.

K: $\Box(\varphi \Rightarrow \psi) \Rightarrow (\Box\varphi \Rightarrow \Box\psi)$
i.e. belief is closed under implication

D: $\Box\varphi \Rightarrow \neg\Box\neg\varphi$
i.e. belief is consistent

4: $\Box\varphi \Rightarrow \Box\Box\varphi$
i.e. the agent believes its beliefs
(termed "positive introspection")

5: $\neg\Box\neg\varphi \Rightarrow \Box\neg\Box\neg\varphi$
i.e. negative introspection

Adding the 'T' axiom to KD45 gives the logic usually referred to as S5, which is commonly used where the accessibility represents 'knowledge'.

T: $\Box\varphi \Rightarrow \varphi$
i.e. what the agent knows is true

N.B., The modal characterisation of mentalistic concepts is the responsibility of the *user*!

Multiple Modalities

The basis of modal logic can be extended to utilise multiple accessibility relations, and hence multiple modalities.

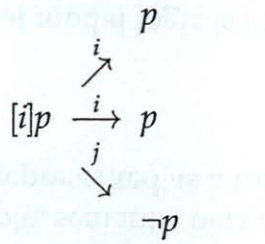
The semantic structure used is now

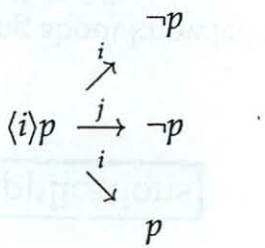
- a set of worlds, W,
- a *set* of labelled binary relations, R_i, on worlds in W, and,
- π, now of type $\pi: W \times \text{PROP} \rightarrow \{T, F\}$

Thus, the syntax of modal logic can be parameterised by the labels i, j, etc:

- '$[i]\varphi$' means that φ is true in *all* worlds that are accessible via R_i from the current world.
- '$\langle i \rangle \varphi$' means that φ is true in *some* world that is accessible via R_i from the current world.

Examples





Common Multiple Combinations

Modalities such as '$[i]$' are often replaced by symbols providing more intuition about the intended meaning of the modality.

For example, if $[i]$ is a modality of the KD45 type, it is often represented by B_i, denoting belief.

If $[i]$ is a modality of the S5 type, it is often represented by K_i, denoting knowledge.

Example 1:
agent i believes that agent j believes φ:

$$B_i B_j \varphi$$

Example 2:
agent i believes that agent j knows φ:

$$B_i K_j \varphi$$

Applications

Practical Reasoning
for example, reasoning about knowledge such as "if I know that you know that...."

Description Logics are the logical basis for conceptual structures (for example, semantic nets and ER models) and can be represented as a form of multi-modal logic.

Program Semantics
for example, a form of modal logic is related to the use of process algebras in representing distributed and concurrent programs.

However.....

Most of the more interesting applications, particularly concerning agent-based systems, require the combination of these modal logics with *temporal* logics.

Temporal Logic: Intuition

Temporal logic is a variety of modal logic where the accessibility relation between worlds is interpreted as a temporal relation.

In the simplest case, all we must do is interpret the basic modalities as follows

'$\Box$' — meaning "in all *future* worlds"

'$\Diamond$' — meaning "in some *future* world"

and then ensure that the accessibility relation mirrors the model of time that we wish to utilise.

Models of Time

There are many different models of time that are used:

linear: each world has at most one future world

branching: a world may have several future worlds

discrete: if $R(w_1, w_2)$ then there is no w_3 such that $R(w_1, w_3)$ and $R(w_3, w_2)$

dense: if $R(w_1, w_2)$ then there *is* a w_3 such that $R(w_1, w_3)$ and $R(w_3, w_2)$

finite past: there is a w_1 such that there is no w_2 such that $R(w_2, w_1)$

The particular models that we will consider are those that are discrete, linear and have finite past.

Discrete Linear Temporal Logic — PTL

The operators used are

- '$\bigcirc\varphi$': φ is true in the next moment in time
- '$\Box\varphi$': φ is true in all future moments
- '$\Diamond\varphi$': φ is true in some future moment
- '$\varphi\,\mathcal{U}\,\psi$': φ is true up until some future moment when ψ is true
- '$\bullet\varphi$': φ is true in the last moment in time
- '$\blacksquare\varphi$': φ is true in all past moments
- '$\blacklozenge\varphi$': φ is true in some past moment
- '$\varphi\,\mathcal{S}\,\psi$': φ has been true since some past moment when ψ was true
- **'start'**: is only true at the beginning of time

Branching Temporal Logic

Note that there are also many varieties of *branching* temporal logics, where the model of time is considered to be a tree branching into the future. Thus, each world may have a number of possible successor worlds.

Examples of Temporal Formulae

Temporal logic allows concise specification of dynamic properties of both individual agents, for example

$$request \Rightarrow reply\,\mathcal{U}\,acknowledgement$$

and the multi-agent system itself, for example

$$broadcast(m) \Rightarrow \forall a \in Agents.\ \Diamond receive(m, a)$$

In addition, formulae such as

$$(\neg passport \vee \neg ticket) \Rightarrow \bigcirc \neg board\_flight$$

constrain the execution steps of the system.

References

Modal Logics

Chellas — "Modal Logic: An Introduction", Cambridge University Press, 1980.

Popkorn — "First Steps in Modal Logic", Cambridge University Press, 1995

Temporal Logics

Emerson — "Temporal and Modal Logic". In *Handbook of Theoretical Computer Science*, Elsevier, 1990.

LOGICAL THEORIES OF RATIONAL AGENTS

Modal and Temporal Combinations

Now that we have looked at (multi-) modal and temporal logics separately, the key element in logical agent theories is the combination of these logics.

For example, a multi-modal logic, on its own, can be used to describe the '*mental state*' of the agent, for example using knowledge and belief.

However, we usually wish to characterise the evolution and change of this state over time – this is where temporal logic comes in.

We first consider two examples showing how such combinations can be generally useful.

Security in Distributed Systems

Temporal logics of knowledge can be used to represent the information that each distributed component is aware of, for example

$$[K_{me}K_{you}key(me) \wedge K_{me}send(me, you, msg)] \Rightarrow \Diamond K_{you}contents(msg)$$

i.e.

if I know that you know my public key, and I know that I have sent you a message, then at some moment in the future you will know the contents of that message

Accident Analysis

$$\left[\begin{array}{l} K_{pilot}\exists engine.\ working(engine) \wedge \\ B_{pilot} broken(left\_engine) \end{array}\right] \Rightarrow \bigcirc shutdown(left\_engine)$$

i.e.

if the pilot knows that there is at least one engine working, and believes that the left engine is broken, then the pilot will shut down the left engine next

Logical Agent Theories

Recall that we said that rational agent theories typically consist of elements relating to dynamic, informational and motivational aspects, and that these are often represented using intentional notions such as belief and desire.

Now we can see which types of modal or temporal logics we might use:

Dynamism — temporal or dynamic logic;

Information — modal logic of belief (KD45) or knowledge (S5);

Motivation — modal logic of goals (KD), intention (KD) or desire (KD).

BDI Agent Theory

The Belief, Desire, Intention (BDI) approach is very popular within agent-based systems.

The theory is built upon combinations of modal and temporal logics, namely:

Dynamism — (linear or branching) temporal logic;

Information — KD45 modal logic of belief;

Motivation — KD modal logic of desire and KD modal logic of intention.

Thus, the temporal logic provides the basic dynamic component, while the KD45 logic allows the agent to have (possibly incorrect) beliefs.

The BDI theory incorporates two motivational elements, which can be characterised as follows:

- desires represent goals that the agent must eventually satisfy;
- intentions represent goals that the agent is *actively* trying to satisfy.

It is the interaction between these motivations that invokes key elements of deliberation.

BDI Example

The behaviour of an agent may be specified in terms of its beliefs, desires and intentions:

$$B_{me} \Diamond D_{you} attack(you, me) \Rightarrow I_{me} \bigcirc attack(me, you)$$

i.e.

if I believe that you desire to attack me, then I intend to attack you at the next moment in time

Alternatively, using just belief and time:

$$B_{me} \Diamond B_{you} attack(you, me) \Rightarrow B_{me} \bigcirc attack(me, you)$$

Problems with Combinations

The combination of (multi-) modal and temporal logics is very powerful.

If there is little interaction between the various temporal and modal dimensions, then such logics are tractable.

However, once we model agents, we tend to introduce many axioms incorporating interactions, for example in a temporal logic of knowledge

(synchrony+) perfect recall: $K_i \bigcirc \varphi \Rightarrow \bigcirc K_i \varphi$

(synchrony+) no learning: $\bigcirc K_i \varphi \Rightarrow K_i \bigcirc \varphi$

As we will see in the next lecture, axioms such as these can make manipulation (for example, proof) much harder.

Summary

- Modal logics are very good for representing the *'mental state'* of an idealised agent, for example characterised in terms of knowledge or belief.
- Temporal logics are very good for capturing the dynamic activity of agents, for example specifying its possible future behaviour.
- The combination of temporal and modal logics give a very expressive framework in which to represent agent activity.
- Many logical theories of rational agents are represented in this way.
- Main problems are:
  - identifying the appropriate modalities to characterise the required agent behaviours;
  - what do we do with such a theory of agency, even when we have defined it?

References

BDI Theory

Rao, Georgeff — "Decision Procedures of BDI Logics". *Journal of Logic and Computation* 8(3), 1998.

Wooldridge — "Reasoning about Rational Agents". MIT Press, 2000.

Combinations of Logics

Fagin, Halpern, Moses and Vardi — "Reasoning About Knowledge". MIT Press, 1996.

Blackburn, de Rijke — "Why Combine Logics?". *Studia Logica* 59, 1997.

DISCUSSION

Rapporteur: Dr B N Rossiter

Professor Dobson wondered what part retribution played in the rational agents structure. Professor Fisher replied that the designer made the rules and decided on appropriate actions. Agents could be destroyed as a punishment if this were appropriate. Professor Sloman thought that the complexities were even more severe than pointed out in the examples with another modal logic needed to do rational planning. Professor Fisher observed that this was why most current systems do not handle planning aspects.

Professor Vogt emphasized that the combination of different logics is a difficult matter and questioned whether a uniform basis was obtained when different axioms were combined. Professor Fisher said that so far they had simply combined all axioms together to produce a common collection. Professor Vogt agreed that you could do this but proof was surely then difficult because of the different basis for the various calculus. Dr Thomsen was also concerned about the consistency and solubility of the logic and thought it would be difficult to verify in derived program constructions. Professor de Marneffe wondered how a programmer would verify the logic in BDI (belief, desire, intention) through identifying actions. Professor Fisher replied that you could add actions as modal logic but it would make the overall logic more complex. Mr Cunningham thought that the practical state of the art in artificial intelligence was to take some concepts from formal philosophy and attempt to use them as design principles. Professor Fisher agreed that there was a big gap in agent technology between current practice and the theory presented in his paper. Current practical systems do not follow the principles of logic.