Technical Report

UCAM-CL-TR-194
ISSN 1476-2986

Number 194

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Proceedings of the First Belief Representation and Agent Architectures Workshop

Edited by Julia Rose Galliers

March 1990

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Proceedings of the First Belief Representation and Agent Architectures Workshop March, 1990

edited by: Julia Rose Galliers
University of Cambridge Computer Laboratory

venue: SRI International Cambridge Research Centre

Abstract

The first Belief Representation and Agent Architectures workshop was organised by Cambridge University Computer Laboratory, and held at SRI International in Cambridge on the 22nd and 23rd March 1990. It was designed as a closed meeting of 15 researchers, all currently working in and familiar with this subfield of AI. The purpose of the meeting was not so much to present completed work, as to to exchange ideas and explore issues with others equally as aware of the relevant problems and background. Each presenter was given 90 minutes in which to lead a discussion on a topic related to their research interests. Generally these were oriented around the presenter's current research projects, outlines of which had been distributed prior to the meeting.

These proceedings comprise eight sections, each including the discussion report followed by copies of the presenter's overheads, followed by the summaries of the presenter's and rapporteur's current research projects. The sections are as follows: General introduction, different styles of agent architectures, a minimalist approach to agent architectures, models of belief revision, the value of formal approaches, knowledge action chance and utility, different value systems, and channels for dialogue.

Contents

Preface

INTRODUCTION

The first Belief Representation and Agent Architectures workshop was organised by Cambridge University Computer Laboratory, and held at SRI International in Cambridge on the 22nd and 23rd March 1990. It was in fact the second workshop of its kind, the first having been the Alvey Workshop on Multiple Agent Systems which took place at Philips Research Labs., Redhill in April 1988.

The nature of these workshops is small, focussed, and discussion oriented. The aim is to facilitate useful interchange between participants about current research problems, issues and ideas.

These proceedings comprise eight sections, each including a discussion report followed by copies of the presenter's overheads, followed by the summaries of the presenter's and rapporteur's current research projects. If the reader would like to contact any of the authors for further discussion or published papers, postal and email addresses follow this preface.

THE SESSIONS

The workshop began with a survey provided by Sam Steel from Essex University. He presented the issues, representational schemes and a list of current researchers involved in the area of the belief representation and agent architectures as related to planning. These and a brief summary by Innes Ferguson from the Cambridge University Computer Lab., comprise the section entitled **Introductory session.**

The next section relates **session 1**, as reported by Colin Hopkins of British Telecom, and presented by Han Reichgelt of Nottingham University. The topic is different styles of agent architectures in which Han compares two competing styles. The first is the more traditionally accepted 'box' architecture in which the agent is designed as vertically decomposed into modules for each capability such as perception, planning and so on. The second is the situated automata model which offers a horizontal decomposition. In the latter, different behaviours as opposed to capabilities are the central issue. Han proposes an architecture which is a blend of the two approaches, currently being developed at the Open University with George Kiss.

Han also briefly describes his current research investigating first-order predicate calculus as an AI knowledge representation language, in this section. Colin Hopkin's described current research is in multiagent planning and the delegation of plans between cooperative agents.

Session 2 comprises a report by Innes Ferguson and presentation by David Connah from Philips Research Labs. David presented his research on a minimalist

approach to agent architecture. His choice of architecture is the second of Han Reichgelt's categories above in which agents are described in terms of their behaviour and are theoretically grounded in situated action. At the workshop, the discussion related to this topic was merged with session 1.

Current research by Innes Ferguson in this section describes a proposed framework for studying the interactions between autonomous plan-forming agents, planning in uncertain environments such as driving along a highway. His focus of interest is plan recognition without unrealistic simplifying assumptions about agents and the domain.

Session 3 reports an overview session about different models of belief revision. The focus is the various problematic aspects associated with incorporating notions of strength of belief to cope with deciding between logically equivalent alternative revisions. The report of the workshop session is provided by Mark Elsom Cooke from the Open University, and the presentation is from Julia Galliers of Cambridge University Computer Lab.

Current research by Julia is described for incorporating reflective capacities related to notions of strength of belief into a belief revision model. A model of autonomous belief revision is a model of choice; choice whether as well as how to revise beliefs, for example during inter-agent communication. Mark is developing computer-based tools for teaching. He describes his aim to build a model of the learner as an active constructor of theories, and how he uses agent design ideas for his dialogue generation and understanding component.

Session 4, entitled 'The Value of Formal Approaches' begins with Nigel Seel's report (STC Technology Ltd.) of the discussion led by Nigel Shadbolt from Nottingham University. This is accompanied by Nigel Shadbolt's slides and summary. The discussion concerns the potential for reconciliation between logicist idealisations of mental phenomena with some psychologically plausible account.

The description of Nigel Shadbolt's current research comprises that of the whole AI group at Nottingham. There is a project on cooperative planning, two on epistemic logics for multi-agent planning systems, and one on planning and instruction. Nigel Seel's current research concerns mathematical models of communication between agents.

Session 5 includes some notes and slides by Sam Steel on knowledge, action, and decision theory. Sam has combined epistemic and dynamic modal logic such that he can represent probability and utility for use in representing decision trees. His interest is in the overlap between decision theory with its emphasis on probable effects and utility of action, and AI planning notions of preconditions and effects. Kave Eshgi from Hewlett Packard Labs was rapporteur for this session. His current research is on the application of model-based diagnosis theory to realistic circuit diagnosis problems.

George Kiss from the Open University presented **session 6** on different value systems. The report is by Jim Doran from Essex University. George presents agents as dynamic entities, described in terms of states and transitions; a behaviour is a trajectory between such state transitions. Goals and values are interpreted within this framework in terms of attractors and repellors.

George's current research investigates HCI dialogue as a special case of action by and between autonomous agents. Focal interest concerns agent attitudes, epistemic, praxiological and axiological, and notions such as commitment and relationship to action. Implementation mechanisms are also being investigated. Jim Doran's contribution concerns a problem for coordination between multiple agents. It relates to vehicles crossing a square, Tiananmen square.

The last **session 7**, as reported by Ann Blandford from the Open University, includes a discussion by Phil Stenton on the work at Hewlett Packard in designing cooperative interfaces to information management systems. The emphasis is on what users want. This contrasts with the other sessions in the workshop in which Phil points out that phrases such as 'principled agent architectures' are used, but he asks what this means? For Hewlett Packard, the principles are clearly user oriented. The issue is grounded in the motivation for work on agency; are we modelling skills or psychology?

Phil's current research involves the analysis of real data from experiments and field study transcripts. The resulting dialogue theories are used in matching interface technologies to dialogue requirements in a financial information system. Ann Blandford is developing a model of tutorial dialogue based on aspects of agent theory.

CONCLUDING REMARKS

This was a most successful workshop. The emphasis on discussion rather than formal presentations was appreciated by all and generally endorsed as the style for next year's follow up workshop. This is to be arranged for spring 1991, and Steve Pulman has generously offered SRI International, Cambridge, to again be the venue. Potential participants can write to me by January 5th 1991, with details of their interests and research. Workshop attendance will be limited to a maximum of 15 participants.

Julia Galliers

Belief Representation and Agent Architectures Workshop

22-23 March 1990

held at
SRI International Cambridge Research Centre
23 Millers Yard, Mill Lane, Cambridge CB2 1RQ
Telephone 0223 324146

Address List

Ann Blandford
IET
Faculty of Social Sciences
Open University
Milton Keynes MK7 6AA
AE_Blandford@uk.ac.ou.acsvax

David Connah
Artificial Intelligence Group
Philips Research Labs
Cross Oak Lane
Redhill RJ1 5HA
connah@prl.philips.co.uk

Mark Elsom-Cooke
IET
Faculty of Social Sciences
Open University
Milton Keynes MK7 6AA
MT_ELSOMCOOK@uk.ac.ou.acsvax

Jim Doran
Dept of Computer Science
University of Essex
Colchester CO4 3SQ
doraj@uk.ac.sx

Kave Eshgi
Hewlett Packard Laboratories
Advanced Management
Information Dept
Filton Road,
Stoke Gifford
Bristol BS12 6QG
ke@hplb.hpl.hp.com

Innes Ferguson
University of Cambridge
Computer Laboratory
New Museums Site
Pembroke Stree
Cambridge CB2 3QG
iaf@uk.ac.cam.cl

Julia Galliers
University of Cambridge
Computer Laboratory
New Museums Site
Pembroke Stree
Cambridge CB2 3QG
jrg@uk.ac.cam.cl

Colin Hopkins
British Telecom Research Labs
Martlesham Heath
Ipswich IP5 7RE
colin@uk.co.bt.hfnet

George Kiss
HCRL Faculty of Social Sciences
Gardiner Building
Open University
Milton Keynes MK7 6AA
best contacted by phone (0908) 652568

Steve Pulman
SRI International
23 Millers Yard
Mill Lane
Cambridge CB2 1RQ
sgp@uk.ac.cam.cl

Han Reichgelt
Dept of Psychology
University of Nottingham
University Park
Nottingham NG7 2RD
han@uk.ac.nott.psyc

Nigel Seel
STC Technology
London Road
Harlow
Essex, CM17 9NA
nrs@uk.co.stc.stl

Nigel Shadbolt
Dept of Psychology
University of Nottingham
University Park
Nottingham NG7 2RD
nrs@uk.ac.nott.psyc

Sam Steel
Dept of Computer Science
University of Essex
Colchester CO4 3SQ
sam@uk.ac.essex

Phil Stenton
Hewlett Packard Laboratories
Advanced Management
Information Dept
Filton Road
Stoke Gifford
Bristol BS12 6QG
sps@com.hp.hpl.hplb

INTRODUCTORY SESSION

PRESENTED BY: Sam Steel

REPORTED BY: Innes Ferguson

Introduction: Areas of Interest

Presenter: Sam Steel

Rapporteur: Innes Ferguson

Sam Steel's introductory talk provided a (very!) fast and condensed survey of issues pertaining to action, knowledge, belief, and planning. The talk was divided into three parts. In the first, a list of relationships was drawn between actions and knowledge, knowledge and plans, beliefs and plans, communication acts and beliefs, communication acts and plans, actions and uncertainty, and learning and planning, among others. Many examples were given to clarify these relationships and these are included in Sam's slides.

The second part of his talk addressed representational issues with beliefs and plans. He listed a number of alternative representation schemes (with examples), together with a brief mention of some of the potential pitfalls and known inadequacies associated with these schemes.

In the final part of his talk Sam gave a (partial) list of those researchers in the field which he considered important. Much to his dismay, he noted that none were Brits and wondered why this was so...

DISCUSSION

During his talk Sam admitted to not being aware of much work relating learning and planning. Nigel Shadbolt contributed various names such as Mitchell and Blythe (CMU) (compiling learnt plans) and Hammond (case-based planning). Hans Reichgelt also felt that work at IBM by Fagin, Halpern, Moses, and Vardi on knowledge in distributed systems was worth adding to the list. Other names mentioned included, among others, Chapman (belief-free agents), Dean et al. (time-dependent planning), and Drabble (qualitative process theory).

Sam also mentioned that if anyone required further clarification regarding the content of his talk that they could contact him directly. His address is provided elsewhere.

INTRODUCTION: AREAS OF INTEREST.

Action & knowledge

Why does belief figure in planning?

(Topics, instead of taxonomy)

2 sorts of connection

- belief as context of planning process
- belief-changing actions as part of plan

1) Most basic:
Knowing how / finding out

— The need of plans for knowledge
- what is Mary's phone number ?
- how does one brail ?

— The effect of action on knowledge
- using telephone directory
- dipping litmus in solution

— how to express actions with
knowledge preconds
effects

after I wash car, it is clean
after I look at car, and it is clean,
then I know it is clean

2) Belief and others

— plan recognition

their plan ⟶
their beliefs
(their values)

— what do I think he will do?
  — does it help me?
  — can I help him?

— what do we think we will do?
  — how represented?
  — how distributed

— Do our belief
or our languages differ?

3) Action & communication I

2 approaches

— My actions (inc. utterances) affect your beliefs

explicit forcible updates

— My actions (inc. utterances) reveal my beliefs & so affect your beliefs

and conversely

your actions / my beliefs

4) Action & communication II

plans as guides to what to say

- I plan to tell you things
  - in this order
  - by these actions

- I talk you on the assumption you have a certain (partial) plan
  - task models
  - student models

- I describe plans to you

5) planning to believe

- plans to know
  - observation
    - selective attention
  - testing of non-observables
    - litmus
  - calculation
    - avoiding logical omniscience
  - recall
    - I dont know X, because recalling X failed (Haas)
- assumption / retraction are actions on beliefs
- the frame problems
  what do I believe when I believe "X is constant unless X is explicitly altered"

NM logic ? NM belief?

6) Action under uncertainty

— expected utility of action A.

$$\sum_i \text{probability of } P_i \cdot \text{utility of } A \text{ if } P_i \text{ holds}$$

— conditional plans

- when are all cases covered?
- need they all be covered?

7) Belief & planning to plan

- dynamic world, so beliefs will change are suspect

- so interleare
  - planning
  - execution
  - perception
    - checking effects
    - general look-out

- when interleaving
  - outcome of each branch is uncertain
  - consider their expected utilities

8) Learning & planning

- what worked last time?
- revise expected utilities of actions after executing them

- Points on representation

— Facts

— modal

B_{tom} dead(fred)

— syntactic

believe(tom, "dead(fred)")

— database (variant of above)

— propositions

believe(tom, p472)

true(p472) ≡ dead(fred)

— paradox lurks

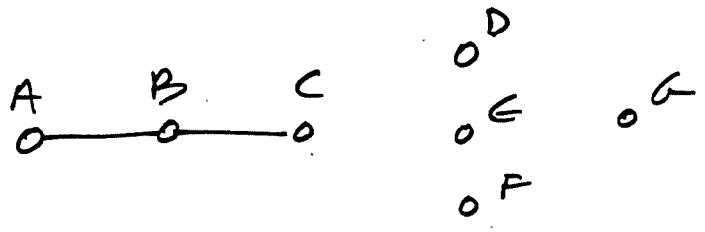
— need for sufficiently fine grain to avoid collapse of denotations

— same issues for terms as for sentences

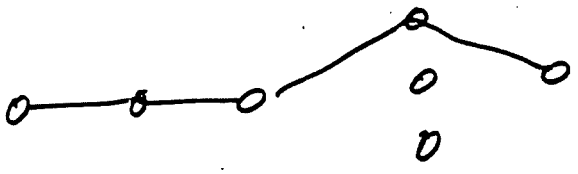
2) Partly-known objects

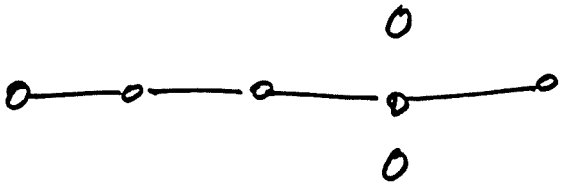
John goes from A to G, via A, B and C.

— as incomplete description



— as alternative complete descriptions

either 

or 

or 

Some names

- Georgeff
  - PRS
  - knowledge-getting tasks
  - stacks of meta & object plans
- Cohen / Levesque
  Logic inc theory of action
- Shoham
  Alternative futures via belief
- Rosenschein / Kaelbling
  - Situated automata
  - K X ≡
    ∀s (s compatible with observation,
    s ⊨ X)
  - strongest preconds & actions

— Russell
expected utility of planning
— Aumann (& game theory)
common knowledge
partitions not relations on frames
— Minton
learning & planning (???)
— Bratman
philosophy of action (???)

All American!

Learning & planning

- learners get interested in plan
- ~~Hutt~~ Mitchell / Blyth:
  Sit. autom & computing
  Learn plans @ MRC (Cambridge)
- case based planning

Knowledge in distrib systems

IBM. restriction on comm ⇒
restriction on whole

Fagin / Halpern / Moses / Vardi
cheating etc

~~Read~~

Reading stuff

Chapman etc. belief-free planning

Brown U.

Dean / Kanazawa / Morgenstern.

Manny Rayner : SICS

IJCAI-89 w/shop

Druddle AIAI

Verification by Ord. sim.

Process Theory

Clarke via CCL. process

SESSION 1:
DIFFERENT STYLES OF AGENT ARCHITECTURE

PRESENTED BY: Han Reichgelt

REPORTED BY: Colin Hopkins

Also reports of current work by:

Han Reichgelt on Agent Architectures

Colin Hopkins on Multiagent Planning

Different Styles of Agent Architecture

Presenter: Han Reichgelt

Rapporteur: Colin Hopkins

Summary of Presentation:

This talk concentrated on a comparison of different ways of specifying agent architectures in terms of the advantages and problems associated with each of them. Three distinct approaches were outlined.

1. The "Box Model" Approach

This approach is based on the vertical decomposition of agent components such a that a set of agent capabilities are identified (planning, perception, belief revision etc.) which are then implemented as a set of separate modules. In such a model, having a propositional attitude (a belief) is associated with having a data structure in the corresponding module.

Two examples of the Box Model approach were discussed. The first was Reichgelt & Kiss' TASTE architecture in which the axiological component of agents has been integrated within Reichgelt and Shadbolt's TEST architecture. The second architecture mentioned was the BDI approach taken by Bratman, Isreal and Pollack which was considerably more complex than the TASTE model in that many more components and interconnections between components were identified.

The principle advantages of the Box Model were threefold:
Modularity - boxes can be brought together as distinct components.
Flexibility - different components can be brought together and tested.
Explicitness - the boxes use explicit data structures, such as beliefs, which gives rise to the possibility of self reflection.

A disadvantage of the box model is associated with its modularity in that boxes are often brought together in a somewhat *ad hoc* manner. What is often lacking is a principled account of the interaction between boxes. A separate disadvantage of the box model arises, perhaps, because of its explicitness, i.e. box models are relatively inefficient. The use of explicit data structures and reasoning can slow box models considerably.

2. The Situated Autonoma (SA) Approach

This approach is based on a horizontal decomposition of agent capabilities such that a set of behaviours that the agent is expected to exhibit (commonly in terms of stimulus-response pairs) are identified and implemented separately. Unlike the box-model approach, having a propositional attitude is identified with being in a state in which an external observer interprets the agent as having a propositional attitude.

A number of examples of the SA approach were cited. The first was that of Rosenschein and Kaelbling (in which the action and perception box are vertically decomposed however). This approach uses the REX language which takes high-level descriptions of behaviours and generates low-level descriptions of hardware that could (in principle) be implemented. The second example cited was Chapman & Agre's PENGI which is a system that constantly examines the world and decides the next action - no actual planning takes place. The third example was the Universal Planning approach adopted by Schoppers which was described as a set of situated actions rules generated for every possible situation. Drummond has pointed out that this may be a non-optimal set and would not cover novel situations. In such circumstances, Nilsson has suggested that a box-model planner would have to be used. Finally, Fagin, Vardii and Moses concern themselves with the problem of attaining knowledge in distributed systems - how it can be said that the system as a whole can be said to have knowledge. The communication protocols between agents will effect that kind of knowledge the system as a whole will obtain. An example (the `muddy child' - a variation of the `cheating husband') was discussed to demonstrate this.

One of the advantages of this approach is speed, agents are not required to `think', just act. Another advantage is that the actions are functionally independent and can be considered in isolation.

Some of the disadvantages of the situated autonoma approach are the obverse of the advantages associated with the box model, i.e. the explicitness and flexibility of the latter is lacking. A second disadvantage of the SA approach is that resource - dependent actions cannot be used (Ginsberg). In situations in which the agent has enough time, it may as well plan its actions. The third disadvantage of the SA approach is that there may be far too many situations that could be captured by SA rules. However, as a counter to this Chapman has pointed out that there is often quite a bit of structure in reality that can be exploited in the use of rules.

3. The Stratified Architecture Approach of Kiss & Reichgelt

This approach, which can be seen as a compromise between the box-model and the SA architectures, identifies the agent in terms of `high level' or abstract capabilities which are then analyzed in terms of lower-level or `primitive' behaviours, cf. programming languages. The agent actions are fast routines which can be combined in novel ways so that new behaviours can be generated.

A set of primitive actions could include adopting an attitude (value, belief or goal), abandoning an attitude and making a choice. A higher-level action that could be composed of such lower-level actions might be deduction. This would be composed of retrieving beliefs, making a deductive inference, making a choice and then adopting the new belief.

The top level of the agent consists of box models which are relatively slow and inefficient but can be easily changed or modified. At this level are explicit, abstract, description of agent

actions which are compiled into situated automaton at different levels in the system which can be run in parallel. As such situated automaton are very fast but difficult to change.

Although one might make the distinction between internal actions (such as deduction) and external actions (such as stacking a block), the important distinction here is between primitive and non-primitive actions. Reflexiveness and learning at the higher-level of the agent are important characteristics of this approach.

There are open questions still to be answered. Firstly, what is the relation between different layers of the model? Secondly, under what circumstances should the bottom layer be re-constructed? Finally, how can beliefs be ascribed to the system as a whole?

Discussion

NS: Why are different levels run in parallel?
HR: Some elements can be changed relatively easily in the box model but not in their compiled form. Therefore, the changing elements can be kept in the box model and time-critical elements can be compiled down for fast running.

SS: I spy a homunculus lurking! How does one decide which actions to compile, which component decides this?
GK: We could regress to different recursive levels of the agent architecture. However, we have a fixed-point agent which is an approximation to a `real' agent. The question remains - where is the agent or its boundaries? The fixed-point is a hint since our agent is and approximation to it.

KE: Rosenschein gives us a grounding in reality from simple logic gates to more abstract agents which keep re-appearing in different ways.
HR: But Rosenschein's SA cannot self reflect - the observer analyses the agent's behaviour and ascribes beliefs to it - the system itself does not self reflect.
KE: There is no reason why, in principle, SA could not ascribe beliefs to itself.

** Discussion stopped here to allow Dave Connagh to give his presentation since it overlapped with HR's (see Innes Fergeson's summary of DC's presentation for details). A joint discussion followed picking up on DC's assertion that SA agents need not hold an explicit representation of their goals.

KE: In order to simulate the world SA need data structures and an algorithm for doing projection and manipulation. A representation is therefore necessary and internal states do no count as models of the world. With such representations humans can do analogical reasoning.

NS: Yes, agents need to distinguish themselves from the world in the same way that babies learn to do very quickly.

DC: By de-coupling the inputs from the world agents can simulate it.

SP: But that counts as a model of the world!

DC: Agents do not have a model in terms of the explicit set of statements that represent the world and is updated by a separate mechanism.

HR: But whatever the agent inspects internally after the simulation is a representation or model.

NS: DC's claim that goals do not need an objective existence needs one question answered; at what point do explicit goals become essential? I would argue that they become essential in forming part of the explanation of the behaviour of other agents.

HR: There is no clear or principled distinction between goals and values. Goals could require more than one action and are the `leaf-nodes' of a value tree. The distinction between goals and values becomes seamless.

GK: There are two questions here - explicit vs implicit goals (I assume DC favours the latter approach) and the hierarchy of agent architectures. As one moves up the agent hierarchy goals become more explicit. The `situatedness' of the agent relates to its causal coupling to the world. As one moves up the hierarchy one finds more de-coupling (less direct causal linking) to and from the world. The usefulness of the higher layers comes from their de-coupling, i.e. reversibility, alternative choices, flexibility, ease of manipulation etc.

DC: Explicit goals become necessary but there is no firm criteria for saying when. Reflection might be one however.

SS: Can values be ascribed to agents in the same way as goals can? They don't actually exist within the system.

KE: What is the distinction and relationship between behaviour and computation?

GK: In the SA approach there would not be a distinction, both would be physical processes. There is a difference between symbolic and non-symbolic activities, the former are maximally de-coupled from the world.

KE: How do we link the abstract description to the explicit behaviours - since we can build a machine to show this?

DC: the two ends of the spectrum, from the box model to the SA approach, could have different assumptions.

SS: Perhaps we don't need goals in the SA but expected utility so that goals can change.

HR: But then there is no difference between utility functions and goals.

Different styles of agent architectures

Han Reichgelt
AI Group
Dept of Psychology
University of Nottingham

"Box" models

Vertical Decomposition (Brooke):
Identify a set of "capabilities" that an agent should have and implement a module for each.

e.g. Planning module
Perception module
Belief module
Axiological module
and so on.

Having a propositional attitude is having a data structure in the corresponding module.

Two examples:

TASTE
Reichgelt and Kiss

The BDI model
Bratman, Israel, Pollack

Advantages:

1. Modularity
2. Flexibility
3. Explicitness
   Possibility of reflection

Problems:

1. Often somewhat *ad hoc*
2. Relatively inefficient

Situated automata

Horizontal decomposition:
Identify a set of behaviours that an agent should exhibit (usually in terms of stimulus-response pairs) and implement each behaviour independently.

Having a propositional attitude is being in a state that an external observer interprets as you having this propositional attitude.

Examples:

Situated automata

Rosenschein and Kaelbling

Pengi

Agre and Chapman)

Universal Planning

Schoppers

Situated action rules

e.g. Drummond

Knowledge in distributed systems

Fagin, Halpern and Moses

Advantages:

Speed

Problems:

1. you do not get the explicitness and flexibility of the "box" models. Hence, no reflection.

2. No resource-dependent action (Ginsberg)

3. There are too many possible situations (Ginsberg) However, there is structure in reality that can be exploited (Chapman).

Primitive agent actions

Identify a set of primitive agent actions and analyze each higher "capability" in terms of these primitive actions.

Cf. Programming languages

The agent actions are fast routines that can be combined in novel ways to generate new behaviours.

Example:

- Adopting an attitude (value, belief or goal)
- Abandoning an attitude
- Making a choice
- Making an inference step
  - perceptual inference step
  - deductive inference step
- Retrieving an attitude
- and so on

Perception:

1. Making perceptual inferences
2. Making a choice
3. Adopting a belief

Deduction:

1. Retrieving beliefs
2. Making deductive inferences
3. Making a choice
4. Adopting a belief

Stratified architectures

Bottom levels consist of situated automata.

Top levels consist of "Box" models.

Agent actions provide an "implementation" language for the "Box" model and can themselves be compiled into a situated automaton.

Different layers running in parallel.

This would combine the advantages of the different approaches.

Open questions:
What is the relation between the different layers?
Under what conditions should the bottom layers be reconstructed?
How can you ascribe beliefs to the system as a whole?

Current research of Han Reichgelt

Han Reichgelt
Psychology Dept., University of Nottingham

My current research in the area of multi-agent systems and planning consists of two strands, namely (i) the (continued) investigation of ways of overcoming the limitations of first-order predicate calculus as an AI knowledge representation language, and (ii) an investigation of different ways of specifying agent architectures, and the advantages and problems associated with each of them.

The first strand of research comprises two subtopics. The first subtopic is based on the observation that there are more expressive logics than first-order predicate calculus, such as modal logics. With Peter Jackson, I have developed new theorem proving techniques for modal logics in general, and epistemic logics in particular. I have recently extended this work so that it can deal with multi-agent epistemic logic. This enables the system to reason about the beliefs of more than one other agent. On two research projects which I jointly hold with Nigel Shadbolt, I am also investigating alternatives to epistemic logic as a formalism for reasoning about belief, and alternative ways of specifying theorem provers for modal epistemic logic, for example, by moving to a reified epistemic logic. This work started a few months ago, and I expect that Nigel Shadbolt will report on this in his talk to the workshop.

A second subtopic under the general heading of limitations of first-order logic is inspired by criticisms such as those of David Israel and more recently Drew McDermott that there are certain important types of reasoning that are not purely deductive. They therefore cannot be adequately dealt with in any system that uses a theorem prover for first-order predicate calculus (or indeed any other logic) as its sole reasoning mechanism. An example is default reasoning. The intuition is that defaults are not implications that can be used in deductive reasoning; rather, they are instructions to the reasoner to add further assumptions about the world to its knowledge base, when the need to do so arises. On making these assumptions, the system can then reason deductively from the enlarged knowledge base. This style of reasoning is called theory extension. I implemented a reasoning architecture that supports theory extension, and used it to built a reasonably sophisticated default reasoner. The advantage of this particular approach is that one does not have to complicate the logic used to represent information about the world. The disadvantage is that one has to complicate the reasoning architecture.

Recently, Nigel Shadbolt and I have applied this framework to planning as well. We argue that there is no purely deductive account for planning either, and that planning should also be seen as involving theory extension. We believe that this analysis gives us both a conceptual framework in which to analyze existing planners, and a specification tool for new planners. The original implementation that was used for default reasoning has been modularized to a greater extent, and we

have tentatively identified 4 heuristics that are used to guide the theory extension process. The resulting system is called TEST (Theory Extension Specification Tool). We are currently investigating whether the four heuristics are sufficient to allow us to express the behaviours of different families of planners.

The second strand of research that is relevant to the workshop is some work that I have been doing with George Kiss on the appropriate way of specifying architectures for artificial agents. In particular, we have compared the vertical decomposition approach in which the overall behaviour of an agent is analyzed in terms of a set of high-level modules, each of which is capable of a particular style of reasoning. So, one may have a planning component, a perception component, a reasoning component, a top-level goal generating component, and so forth. An alternative approach is based on horizontal decomposition in which an agent is analyzed as a set of behaviours that it needs to exhibit in specific situations. The idea is to give a set of situated action rules, and to ensure that the agent exhibits the correct behaviour in each situation. Rosenschein's situated automata approach is perhaps the best example of this. A final approach that George Kiss and I have looked at is a compromise between the two. We analyze agent behaviours in terms of a list of primitive agent actions. These actions are higher-level than the action considered by the situated automata people, but not as high-level as the vertical decomposers'. An example would be the primitive action of adopting a belief. The idea is to analyze more higher-level agent actions, such as drawing a deductive inference, in terms of these primitive actions. The extent to which we are able to do so is an open question. I will discuss these issues in more detail in my presentation to the workshop.

Current research of Colin Hopkins: Multiagent Planning

Colin Hopkins
British Telecom Research Labs., Ipswich IP5 7RE

Enabling planning systems, or 'agents', to operate in a cooperative manner is the central aim of my present research. In doing so the main focus of interest is the construction of plans in which the planning responsibility for some of the sub-goals can be delegated to other similar agents. As part of this research I have developed a computer-based multiagent planning system, called 'DePlan', which can distribute planning goals to other agents by *DE*legating the *PLAN*ning of those goals to those agents. A more detailed account of DePlan can be found in Hopkins 1988 and Hopkins 1989.

Constructing such 'multiagent' plans requires beliefs to be held by the computer agent about its world and the inhabitants of it. However, there is currently a trade-off between resources devoted to belief modelling and derivation and the more general process of problem solving. Because DePlan is concerned with the latter, a relatively simple belief representation framework is used. DePlan's real advances must be seen in the light of its action modelling, its communication protocols and its overall planning framework.

Action Modelling

In addition to beliefs, agents may attempt to model, within their own plans, the goals that may be achieved by other agents (in terms of the actions that the other agents may execute in order to achieve those goals). This is seen as desirable (although not strictly necessary) since it allows an agent which is producing a plan to include actions executed by other agents so that the effects of these actions can be relied upon within the former's plan.

In previous research (e.g. Konolige and Nilsson 1980, Corkill 1979), it has been considered sufficient to model action execution within an agent's plan. However, the range of actions that could be executed by other agents may be so complex that they amount to plans in themselves. This requires the planning agent to reason about the planning behaviour of other agents during the construction of its plan. The process of reasoning about planning as part of one's planning process is, of course, meta-planning and so the meta-actions of plan elaboration and execution are represented within an agent's domain level plan. Mixing such operators is derived from 'Cross Level Planning' (Bartle 1988).

Communication

Having modelled the cooperative action of other agents within one's plan, the next step is to 'motivate' the designated agent to perform the required action. This is because an agent must have some reason to believe that the other will actually perform the desired actions. Requests are an obvious solution here. Requests, however, can vary in range and complexity in that a request itself may not be a single

executable act but one which may involve a certain amount of planning. An agent may request not only single goals to be achieved but a whole set of actions to be executed that will achieve a desired goal. In effect, what one is requesting at this end of the spectrum is a whole plan that is to be executed. Examples of where this may be important can be found in cases where one is passing a complete plan as a detailed set of instructions to a person who may not be entirely competent at a particular task or where the requestor is concerned that the goal is achieved in a particular way (so as to avoid undoing previously achieved or desired goals for instance). Since plan delegation is a technique that can be used to produce requests ranging in complexity from simple goal achievement to plan execution, it can be seen to subsume previous formulations of requests for goal achievement.

Planning

DePlan can be viewed as an extension of existing 'classical' AI planning techniques. The term 'classical' refers here to those planners which are descendants of STRIPS, the historical antecedent of many of the most well known automated problem solvers such as NOAH (Sacerdoti, 1977) and NONLIN (Tate, 1976). DePlan itself makes direct use of IPEM (Ambros-Ingerson, 1987) which, in addition to sharing the characteristics of these post-STRIPS planners, is able to interleave the process of plan construction (elaboration) with that of plan execution.

Both modelling actions and requests have to be incorporated within the agent's plan. In doing so the agent must make decisions at many levels of planning activity. Specifically, agents are concerned with the sequencing of particular operators to be incorporated into the plan. This plan construction process takes place as a cycle of activity over two stages. At the first, or Task Allocation level, a high-level 'strategic' plan is built up representing sub-goals and the agents responsible for achieving them. This plan then controls the construction of a second 'Flaw-Fix' level in which a 'tactical' plan is built up that details agents to particular tasks for achieving sub-goals and directs the way in which the final plan is to be constructed. At this second level communication operators are scheduled into the plan in order to make agents aware that actions are required of them. At both levels of activity, the resulting 'plan' is, in fact, a schedule of outstanding planning tasks to be completed along with available actions which achieve those tasks. The result of these levels of planning is that DePlan agents construct a nonlinear hierarchical plan in which the planning behaviour of other agents is incorporated in order to achieve some of the sub-goals contained within that plan.

Postponing Planning

Agents can utilize the same process of delegation in order to postpone the planning necessary for their own sub-goals. The rationale underlying this is the fact that modelling the future actions of oneself is no different from modelling the future actions of other agents, except that the agent to which the task is delegated is oneself at some future time. This would seem quite natural given the fact that there seems to be no reason why an agent cannot view its future self, and the future plans it intends to work on, as objects in its domain.

Conclusions

This article has briefly described DePlan, a multiagent planning system that attempts to solve its problems by cooperating with other similar systems. The central technique underlying DePlan's cooperative

behaviour is the delegation of sub-goals in the form requests for action to be taken on partially elaborated plans. The planning behaviour undertaken in response to such requests is incorporated into an agent's plan as actions which model such behaviour. DePlan's planning framework is particularly flexible in that it allows a range of requests to be made, from requests for the achievement of simple goals to requests for complex action to be taken on plans.

References

Ambros-Ingerson J. (1987), *IPEM: Integrated Planning and Execution Monitoring*, M.Phil. Thesis, Department of Computer Science, University of Essex, Colchester, U.K.

Bartle, R., (1988), *Cross-Level Planning*, Ph.D. Thesis, Dep. of Computer Science, University of Essex, Colchester, U.K.

Corkill, D, (1979), *Hierarchical Planning in a Distributed Environment*, Technical Report no. 12, (also in proc. 6th IJCAI, pages 168 - 175) Department of Computing, University of Cambridge, Cambridge, Mass., U.S.A.

Fikes R. & Nilsson N. (1971), *STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving*, Technical Note 43, SRI International, U.S.A.

Hopkins, C. (1988), "DePlan: Enabling Agents to Produce Plans that Achieve Cooperative Problem Solving", *Proceedings of the European Conference on Artificial Intelligence*, Munich, pages 421 - 426

Hopkins, C. (1989), "Meta-Level Planning for Multiagent Cooperation" in *Proceedings for the Conference for the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, pages 185-190, Pitman Press, UK.

Konolige K., Nilsson N. (1980), "Multiple-Agent Planning Systems" in *Proceedings of 1st National Conference on Artificial Intelligence*, Stanford, CA, pages 138-142 American Association for Artificial Intelligence

Sacerdoti E., (1977), *A Structure for Plans and Behaviour*, Elsevier North-Holland

Tate, A. (1976), *Project Planning Using a Hierarchic Non-Linear Planner*, DAI Research Report No. 25, DAI, University of Edinburgh, Edinburgh,

SESSION 2:
A MINIMALIST APPROACH TO AGENT ARCHITECTURES

PRESENTED BY: David Connah

REPORTED BY: Innes Ferguson

Also reports of current work by:

David Connah on multiple agent systems

Innes Ferguson on 'Touring Machines': Rational planners in open worlds

A Minimalist Approach to Agent Architectures

Presenter: David Connah

Rapporteur: Innes Ferguson

David Connah started by declaring that he wanted to avoid defining the meaning of 'minimalist', hoping instead that a meaning would emerge in the course of his talk. The main aim of his talk was to describe and justify three principled, architectural features of agents in real-world settings: situated action rules, schemas, and simulation. The term 'principled' here alludes to David's concern with embedding structures in agents that are a response to certain broad, generic agent needs, rather than imposed as a result of trying to implement some arbitrary block diagram.

Before describing the proposed agent architecture, a brief description of the agent's world was given. In order to avoid placing premature, arbitrary bounds on the scenarios under consideration, the agent's world would be open, dynamic, and populated by multiple agents. A point made, however, was that although this world was dynamic, it would, from an agent's point of view, be locally predictable over short time scales.

David rejected the notion that an agent's activities could be adequately and plausibly expressed in terms of specialized, communicating modules within the agent. Instead, he proposes a theory of activity that is a variant of the so-called layered theory of agent activity. This new theory describes agents' activities in terms of situated action rules, schemas, and simulation.

Situated action rules are considered adequate for describing the types of fast, unthinking action that agents perform on a very frequent basis. The ability to act situationally was seen as being required not just for survival, but also in order to perform many everyday activities, in particular, those which occur over short timescales. At this level, agents store no internal model of the world, but rather use the world itself as its only representation.

In order to manage activity over longer and more varied timescales, as well as being able to cope with routine activity (e.g. making a cup of tea or going to work), the notion of schemas was proposed. Schemas essentially prepare an agent to receive certain kinds of information and thus control the agent's acitivity of seeking this information. It is expected that many such schemas would be operating within an agent, each controlling activity on different timescales. A consequence of this is that there will often be conflicts between the actions required for different schemas. The resultant activity of the agent is determined by a style of constraint satisfaction, rather than by a serial interleaving of possible actions. An obvious question to ask at this stage is how goals get represented; in particular, should goals have an objective existence in the agent? David argues that they

should not. Instead, goals should merely be treated as predispositions to behave in a certain way in certain circumstances. (Note that agents are allowed to ascribe objective goals to other agents, but these are only external descriptions of others' goals.)

Finally, in order to deal with certain conflicts arising from unanticipated schema interactions, as well as being able to operate in novel situations, agents will require some mechanism for viewing the outcome of situations before the situations are actually realised. David calls this mechanism simulation, and believes it can be implemented using situated action rules and schemas (appropriately modified to inhibit the effecting of actions, and in some cases, to inhibit sensory input as well). David noted that how and when agents would employ or control this mechanism was still a subject of research.

David concluded by defining 'minimalisim' as employing a small number of very fundamental processes to create the rather complex overt behaviour of an agent. In his talk he proposed three candidates for the bottom levels of the agent architecture and was confident that while subsequent layers might be built on top of these, many of the desirable properties of his real-world agents would be emergent from the mechanisms he had described.

DISCUSSION

The first topic related to whether or not the act of simulating required an agent to store a model of its world. This in turn raised the questions of (i) whether this model would be explicitly or implicitly defined, and (ii) how an agent might distinguish between the real world and its simulated world. David argued that a world model may be required at some higher level, but that it needn't be explicitly defined or independently controlled and updated in the three layers of his architecture.

The next issue discussed was whether or not some or all of an agent's goals need be explicitly represented. George Kiss noted that the distinction between actions and goals is, in general, unclear, and that the degree of coupling between causal input and an agent formed a continuum of goal abstraction starting with the situated action view and extending up to the plan-based view. George argued that an agent needs to abstract in order to perform "richer" functions. David agreed that functions such as reflection and goal ascription might require goals to be explicitly represented, but how such goals might be represented was not clear.

Finally, there were some general concerns about how one could bridge the gap between the bottom-up, situated action view of agents, and the top-down, plan-forming view. As Nigel Seel commented, the levels at which agents can be described are critical, and the important thing is to relate these different levels rather than attempt to reduce them to some single, primitive level. The discussion tied in strongly with Hans Reichgelt's presentation which introduced the notion of primitive agent actions.

Belief Representation and Agent Architectures Workshop
Cambridge March 1990

A Minimalist Approach to Agent Architectures

David Connah

Artificial Intelligence Group
Philips Research Laboratories
Redhill, Surrey, UK, RH1 5HA

The World

The chief characteristics of the world are

it is open

it is dynamic

it contains other agents

It is also true that the world is complex and dangerous and that things can happen quickly in it.

The Agent

The sort of agent I want to focus on is a domestic robot. Let us call it Jeeves. It has two in–built concerns:

it must try to survive

it is designed to serve

The latter concern implies

the ability to perform certain tasks (its *raison d'être)*

the ability to communicate, cooperate, negotiate.

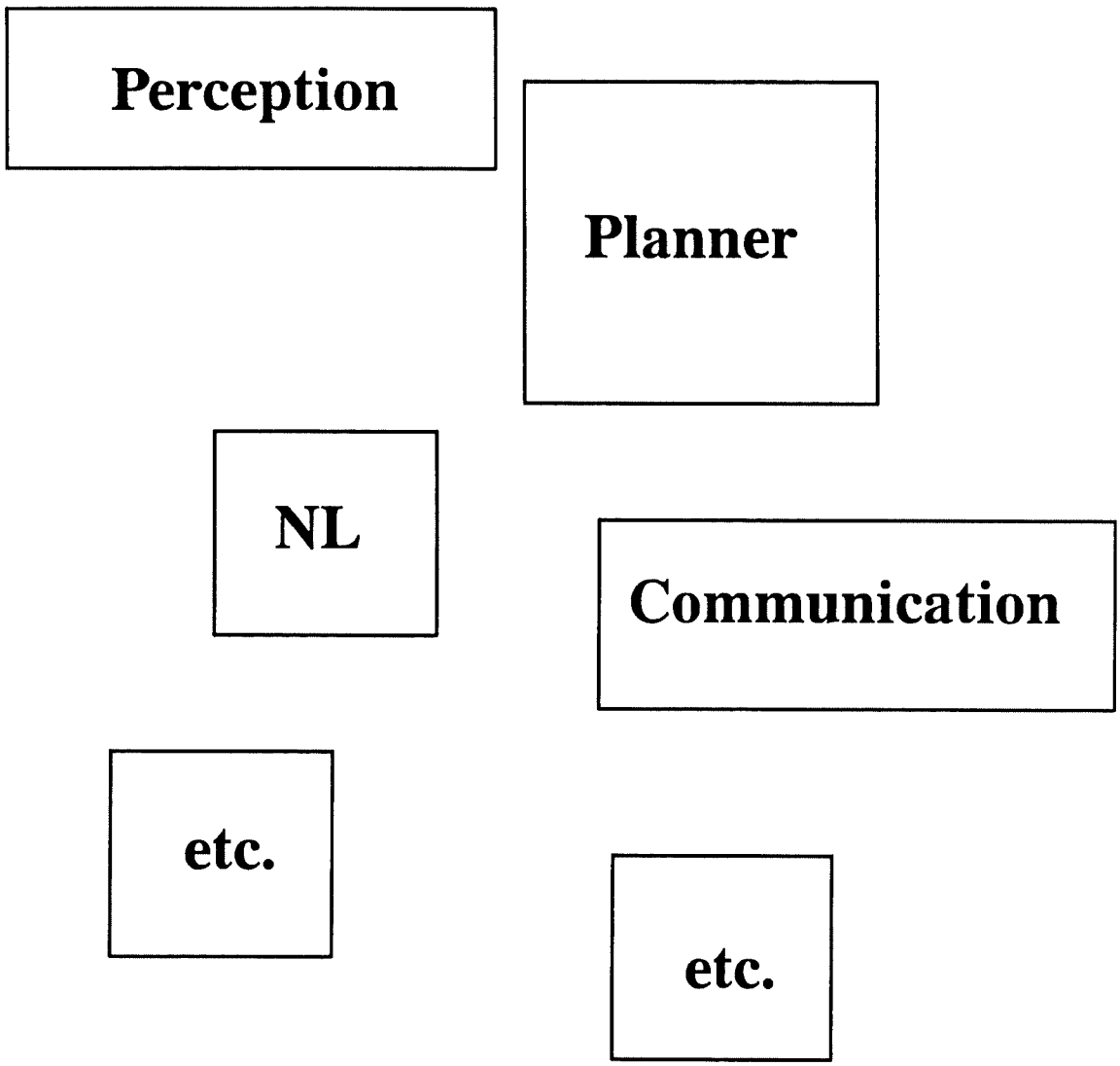
Theories of activity

Two broad types of theory

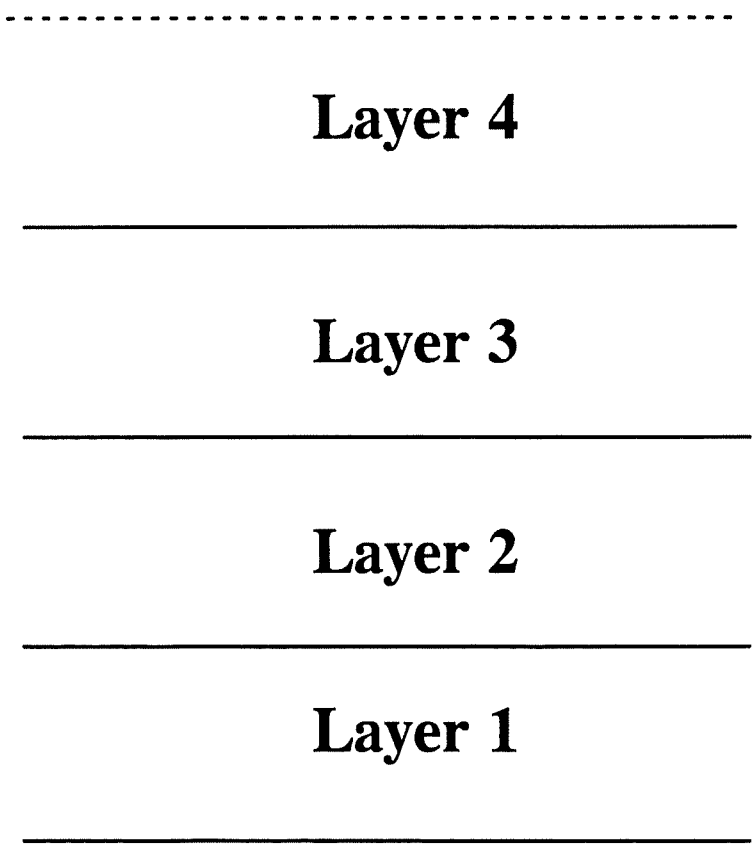
Modular theories

Layered theories

Modular Theories



Layered Theories



Situated Action

Features of situated action

quick, unthinking, action

necessary, in general, for survival

also useful for many everyday activities

(short time scale, direct reference to the environment)

Examples of Situated Action – 1

"A Robot that Walks; Emergent Behaviour from a Carefully Evolved Network." Rodney Brooks.

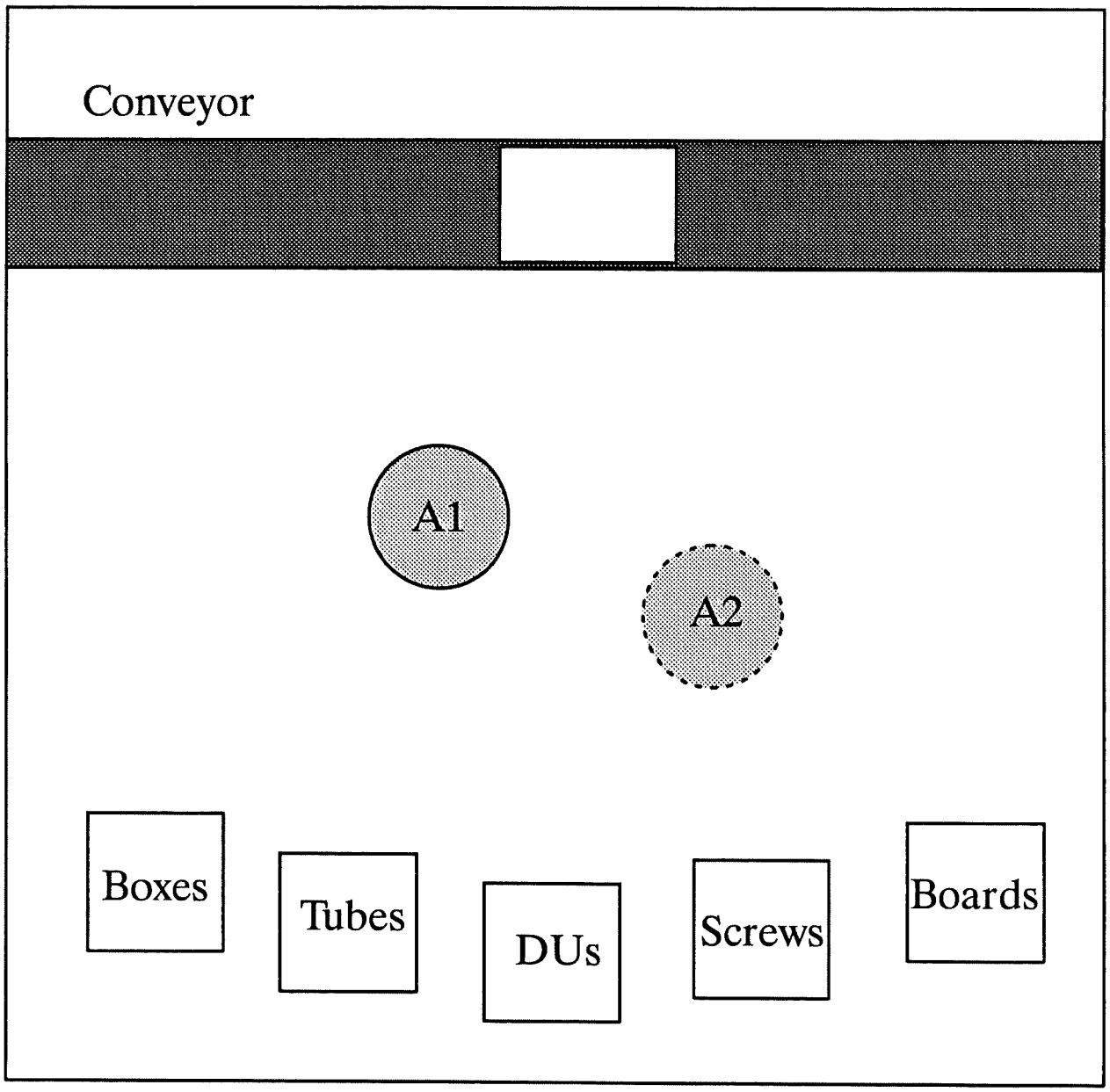
Concerns a six–legged 'insect' which has two motors per leg – back/forward and up/down.

The Behaviours

1	**Stand Up**	stands up when powered up
2	**Simple Walk**	one leg forward – the rest back (gait)
3	**Force Balancing**	compensate for rough terrain
4	**Leg Lifting**	trade–off speed/obstruction height
5	**Whiskers**	feelers
6	**Pitch Stabilization**	compensate for pitch instability
7	**Prowling**	tends to follow IR source
8	**Steered Prowling**	takes account of general direction of IR

Examples of Situated Action – 2

Television Assembly



Schemas

(schemata if you prefer, I shall stick to schemas)

Schemas are designed to do two things for us

1 Cope with longer time scales than situated action

2 Handle routine activity

Schemas prepare the agent to receive certain kinds of information and thus control the activity of seeking this information

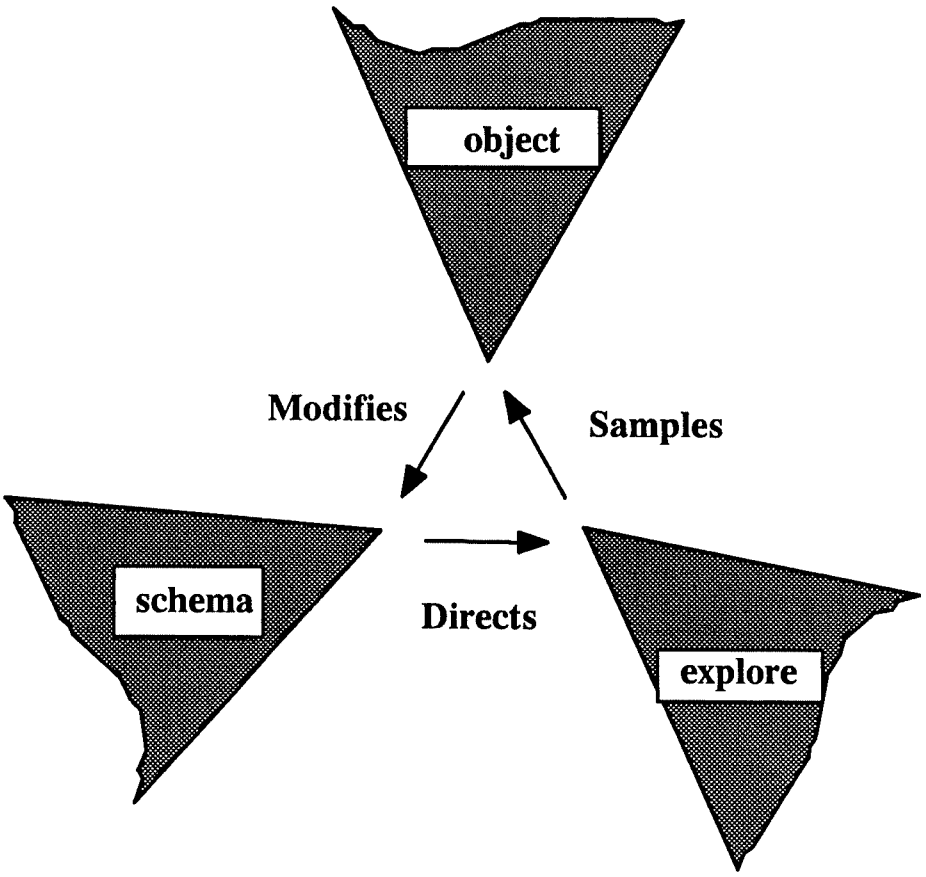
Examples of activities that might be covered by schemas:

tying a shoelace

making a cup of tea

going to work

Perceptual Cycle



Resultant Activity

One can think of the *resultant* activity of an agent in terms of two ideas:

Concurrency

Constraint Satisfaction

Many schemas (some of them requiring situated action) are active concurrently. The actions which the agent performs are the result of constraints at various levels placed on the actions called for by these schemas.

Language can also be thought of as fitting into this pattern of constraints:

The effect of an utterance is to modify the behaviour or the understanding of another agent by systematiclly putting constraints on its situation.

Viewed in this way language is just one of many constraints on the behaviour of an agent.

Goals

Do (or should) goals have an objective existence in the sort of agent we are describing?

My (tentative) answer to that is no. The nearest I can get to a definition of the kind of goal I am talking about is

'a disposition, on the part of the agent, to behave in a certain way in given circumstances'

Examples: (some possibly contentious)

isolated bacteria forming structured colonies for some purposes

termites building nests

getting to work in the morning

making a cup of tea

Simulation

Problems for Situated Action and Schemas:

conflicts

novel situations

Simulation involves the use of many of the normal mechanisms of an agent (e.g. situated actions and schemas) but without the usual accompanying actions and, sometimes, without sensory input either.

Simulation is concerned with anticipation. It can form the basis of *reasoned* behaviour and of *understanding*.

Minimalism

A comparatively small number of very fundamental processes underlies the rather complex overt behaviour of agents. What is required when considering the architecture of agents is that these processes should be defined and implemented. The processes should be chosen in the context of an appropriate theory of activity.

The implication is that we mustn't *directly* implement everything that we can ascribe to an agent.

Current research of David Connah: Multiple Agent Systems

David Connah
Philips Research Labs., Redhill RJ1 5HA

Aims

We are primarily interested in the architecture of autonomous agents and of their interactions, particularly in terms of the cooperation, competition and negotiation between them. In general we expect such agents to interact heterarchically rather than hierarchically.

Approach

We have specifically rejected the idea of treating agents as being primarily involved in problem solving, preferring to start from the assumption that agents are embedded in their environment and have to be able to cope with whatever happens in that environment on an ongoing basis. Higher level activities have to be grounded in that aspect of the agent's architecture which allows this. It seems to us that this approach is helped by choosing to describe agents and the world in terms of behaviour rather than of knowledge. Apart from any other considerations it makes it easier to describe agents, objects and relationships (laws) in a uniform way. Furthermore in a multi-agent situation where each agent may have to observe and interpret the actions of other agents the visibility of much behaviour may be an important factor.

Another decision which we have taken and which we think is consistent with the above approach is to base the bottom level of our agents on a theory of situated action (Suchman 1987, Chapman and Agre 1987). Amongst other things this decision considerably alleviates the problems of intractability associated with other approaches particularly when taken in conjunction with the importance that we attach to the matter of focus of attention. Another aspect of this paradigm is that it sets the role of internal models or representations in quite a different light; in particular, in many situations, no internal representation is needed. We also need to define what terms such as goal, belief and those describing other intentional attitudes mean in the context of this kind of agent.

This decision has occasionally given rise to misunderstandings about what our agents can and can not do. We should emphasise that this situated action foundation is what controls our agents at the lowest level or the finest grain size. Typically such actions will occur only over very short time intervals. Higher level activities will not be directly implemented in this way but will nevertheless still be grounded in situated action. There has never been any suggestion that *all* the actions of an agent are direct reactions to its environment. This raises another point: there is a strong suggestion that what we have said and done so far in the project implies the concept of *emergence*. I think that this is something we are going to become increasingly interested in although we have not done much on this so far (Steels 1989).

Status

We have developed a language (ABLE) and a software tool (LYDIA) which we use for experiments on the architecture and interactions of agents.

ABLE is fundamentally a language for describing or specifying behaviour. There is an interpreter in LYDIA which allows such specifications to be 'run'. The result of interpreting ABLE text can be viewed as a simulation of the total system specified. LYDIA also contains tools which allow the user to debug

the code, to retain and analyse histories of simulations and to ask questions about why certain things did, or did not, happen during a simulation. Using ABLE and LYDIA we have written a number of demonstrations both to test the software and to try out some of our preliminary ideas about agents and their architecture (Hickman and Shiels 1990). One thing that was apparent from these experiments was that the system was very slow. An important strand of our work at present is in speeding up the interpretation of ABLE and, to some extent, in cleaning up the syntax and the operational semantics of the language.

Future work

We have three other principal concerns at present. We are exploring what is the best way of understanding the 'meaning' of an ABLE text (this is not the same as the declarative semantics (Connah and Wavish 1990)), we are beginning to look at the next step in extending the architecture of our agents and we are considering in more detail what sort of problems will be raised by application areas of interest

It is too soon to report on where the first two parts of this programme are leading us but a few words about scenarios might be of interest. We have considered a number of scenarios to try to find one which might be a useful vehicle for our research and at present we are planning to use the smart house for this purpose. There are many interesting problems concerned with the interaction between devices in the house or between the occupants and those devices. For example how does the house know what are the intentions of the occupant at a given time? Clearly he/she can tell it but there may be more subtle ways of finding out. How can added functionalities such as preserving the security of the house and its occupants emerge from the more primitive behaviours of cameras, telephones, monitors etc.? How can the conflicting requirements of the different occupants be simultaneously satisfied perhaps by some compromise and how is this compromise reached? What extra functionalities can be achieved simply through the intelligent cooperation of devices within the house?

References

Agre and Chapman 1987

Agre, P.E. and Chapman, D., 'Pengi: An Implementation of a Theory of Activity', Proceedings of the AAAI Conference, Seattle, Washington.

Connah and Wavish 1990

Connah, D.M., and Wavish, P.R., 'A language for describing behaviour' to be submitted to ECAI90.

Hickman and Shiels 1990

Hickman, S.J. and Shiels, M.A., 'Situated Action as a Basis for Cooperation', to be submitted to ECAI90.

Steels 1989

Steels, L., 'Cooperation between distributed agents through self-organisation' AI Memo No. 89-5 AI Laboratory Vrije Universiteit Brussel

Suchman 1987

Suchman, L., 'Plans and Situated Actions: the problem of human machine communication', Cambridge University Press.

Current research of Innes Ferguson: 'Touring Machines': Rational Planners in Open Worlds[1]

Innes Ferguson
University of Cambridge, Computer Lab.

Plan Recognition on the Highway

When planning in the real world – under uncertainty and in the presence of multiple agents – a planner's abilities to reason about other agents' actions and plans and to understand the causes behind any existing uncertainty will be of vital importance. Recognizing plans is a useful endeavour since it can lead the planner to a better understanding of agents' observed and anticipated behaviour. This in turn will enable the planner to interact more effectively with each agent, as well as improve its ability to predict (and subsequently resolve) likely goal conflicts. Plan recognition is, however, a complex and inherently defeasible task since, in general, there will exist several ambiguous ways to interpret any sequence of actions an agent might perform.

My primary interests lie in empirically studying the interactions between autonomous, plan-forming agents in uncertain environments, an example of which is highway driving. In such a domain, agents – drivers – will be self-interested with respect to their own goals; at the same time, however, they must remain attentive to what is going on around them since certain external events will from time to time affect their own goal-related activities. Since agents will have a limited view of the world, no global goal to work toward, and little or no a priori knowledge of each other's beliefs and intentions, inconsistencies between agents – and thus uncertainty – will arise.

While much progress has been made in the area of plan recognition over the past decade, much of the success of previous approaches rested on making certain important simplifying assumptions concerning the make-up of agents and the characteristics of the domains in which these agents operated. One of the aims of my research is to address which of these assumptions can *realistically* be held, and which, given the very nature of open environments (in this case, highway driving), should be dropped and handled directly by the planner. One of the most common assumptions made in the past, for instance, is that domains contain at most one actor – or planner – and one observer. In such domains, the (passive) observer's unique role is to recognize the actor's plans; the goal-seeking actor, on the other hand, need not concern itself with the observer at all. In the highway domain, where each actor is itself an observer (i.e. where there are several planners), attention will have to be paid to interactions – harmful or otherwise – between actors' plans. A second assumption generally made is that the observer has fairly complete and/or correct knowledge about its world. For example, the observer might be given a complete plan library describing all of the actor's possible actions, or similarly, the observer might be provided with correct beliefs about the (sole) actor's intentions. Given the unpredictable and inconsistent nature of the highway domain, it is unrealistic to assume that every agent will be fully knowledgeable about every other agent's actions – particularly at the level of detail required to infer complete plans.

---

[1]This work is supported by a Bell-Northern Research Ltd. Postgraduate Scholarship and a UK Overseas Research Student Award.

A Proposed Framework

Dispensing with these two assumptions has major implications on the responsibilities of individual agents and a major effect, consequently, on the design of a framework for plan recognition. My proposed architecture is illustrated below. As can be seen, the architecture is intended to enable the agent-planner or 'Touring Machine' to perform three major functions: spatio-temporal planning, evidential reasoning, and world and agent modelling.

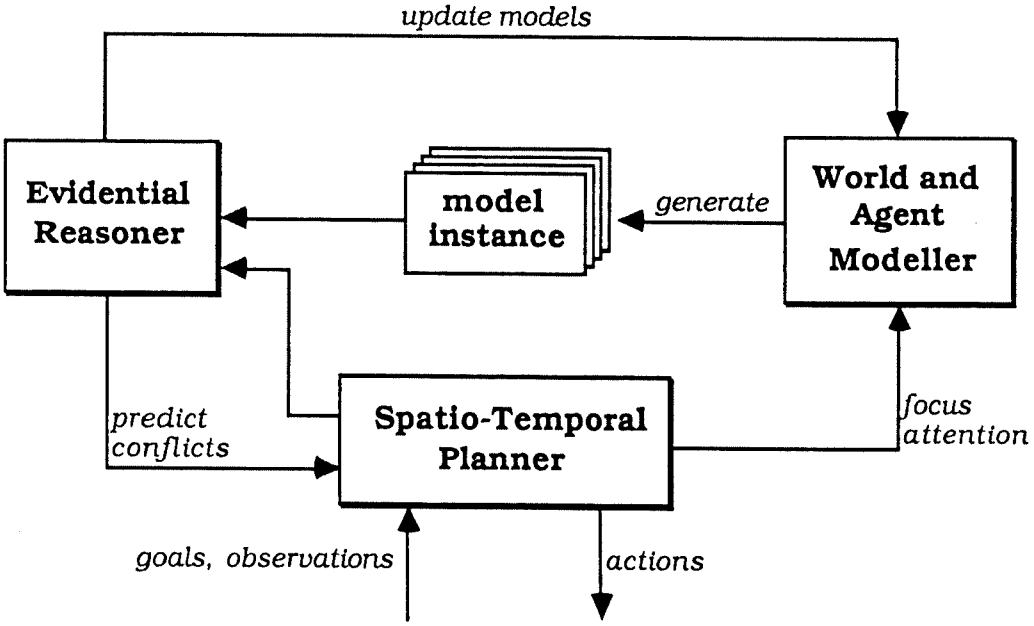


Figure: Plan Recognition Framework

Sensory information and the agent's own actions and goals are processed by the Spatio-Temporal Planner (**STP**). Besides incorporating a temporal logic for constraint-directed planning, this module is responsible for providihg the World and Agent Modeller (**WAM**) with the agent's current focus of attention (e.g. critical agents to consider, newly encountered hazards). The role of the **WAM** is to contruct models of all other agents and the world as perceived by the Touring Machine in question. Influenced by Bratman's action/intention/plan framework of rational agenthood, the **WAM** provides the Evidential Reasoner (**ER**) with inferred accounts of agents' current and projected beliefs, desires, and intentions. Armed with these inferred models, domain heuristics, evidence-driven uncertainty handling procedures, as well as information concerning the agent's own goals, the major tasks of the **ER** will be to inform the **STP** of any predicted spatio-temporal conflicts between agents' plans, as well as providing the **WAM** with appropriate updates to incomplete or incorrect world/agent models.

The proposed framework is expected to provide rational planners with the extra level of robustness and autonomy which is required when operating in an uncertain domain such as highway driving. Through the integration of the techniques illustrated above it is anticipated that there will be no need to make the two aforementioned assumptions. The framework will be evaluated in a simulated world of Touring Machines driving on the highway. The primary aim of the experiments will be to investigate the patterns of interaction that unfold between multiple agents, as each attempts to accomplish its own

goals while coping with the dynamics of the world in which it is operating. By varying agents' reasoning skills, intentions, tolerance to uncertainty, sensing horizons – in short, their 'identities' – insight should be gained into the true potential of the various ideas and techniques being prescribed.

Work is under way on a scenario-driven simulation environment with which to study the behaviour of Touring Machines under different agent and world conditions. The major challenge remains to combine principles derived from Bratman's framework of rational agenthood – in particular, Cohen & Levesque's notion of persistence and Gallier's property of agent preference – with both James Allen's action/time formalism and Paul Cohen's theory of endorsements. Previous work by Georgeff and Pollack in applying Bratman's principles in mobile robot planning domains is also being considered.

SESSION 3:
MODELS OF BELIEF REVISION

PRESENTED BY: Julia Galliers

REPORTED BY: Mark Elsom-Cooke

Also reports of current work by:

Julia Galliers on autonomous belief revision

Mark Elsom-Cooke on educational agents in teaching systems

Models of Belief Revision

Presenter:Julia Galliers

Rapporteur: Mark Elsom-Cooke

The focus of this talk was on the way in which belief revision can be handled in an agent. The basic assumption is that belief revision corresponds to changes in cognitive state, and that these can be modelled using non-monotonic reasoning methods. The particular problem considered was that of deciding when to modify one of a number of beliefs. Work in this field tends to assume that beliefs are either justified or not justified, but gives little support to thinking about strengths of justification. This leads onto the problem of how to determine relative strengths of beliefs and justifications, which can be regarded as a problem of imposing a partial ordering on the elements of a belief set.

1. Predetermined ordering vs end-product of reasoning:
   One approach to ordering is to apply a predetermined strength to fundamental beliefs and to combine these in order to produce a numerical strength for any beliefs justified using that belief. This is somewhat ad hoc in nature, although models of *weights of evidence* can be applied. A second possibility is to reason about the sort of justifications which a belief has. For example, beliefs which have been produced using a default reasoning strategy are weaker than beliefs created from a specific, explicit strategy. This has the advantage of giving a more principled method for talking about the justifications of beliefs and explaining why one is preferred to another.

2. Nature of belief ordering:
   There is a difference between having beliefs which have a particular strength (or certainty) and beliefs which are held, but have varying difficulty of revision. This latter model implies some idea of persistence of belief, and it is the strength of this persistence which will determine when a belief can be changed. This has a relation to the notion of the amount of work involved in modifying a belief, and can be related to the idea of *utility* of a belief within a system or a particular context.

3. Representations of ordering: qualitative vs. quantitative:
   The idea of having a strict quantitative model of belief is appealing in that it would allow a system to guarantee to make a decision. Unfortunately, this approach suffers from a lack of adequacy of such models. It is not clear that the sort of phenomena modelled in belief systems are amenable to a complete representation of their ordering in the way that would be required for a quantitative model. Qualitative representations, on the other hand, could provide a less complete, but more tractable view of the ordering. Such a

representation has the advantage that not only can the uncertainty of a belief be used in reasoning, but it is also possible to reason about the uncertainties themselves.

4. Theories of rational belief change:
   A distinction was made between two basic models of when belief change occurs. In one case, a belief can be held as long as it has a justification (or is a self-evident belief). If that justification ceases to exist then the belief can no longer be held. The alternative is to support beliefs unless a positive reason is found for giving them up. This results in very different properties for the system. The former approach (as exemplified in ATMS) constitutes a foundational theory. There are fundamental, unchangeable beliefs upon which everything else is built. In the latter approach it is possible to construct sets of beliefs which, because they provide a coherence, can be regarded as being independent of a fundamental belief (if necessary). These beliefs form mutually supporting groups which continue to exist while they serve a role in the system. Such groupings cannot exist in foundational systems, because there must be at least one belief which is supported only because there is no evidence to the contrary.

5. Modelling context:
   A final issue to be raised was the way in which the sort of belief revision which we support must be related to the goals of the modelling which we perform. The formal approaches, while providing guaranteed properties of consistency etc. are not of a form which is currently computationally tractable. For *real* problems it seems likely that we will wish to use less completely specified informal modelling methods which are computationally tractable, but which lack a proper semantic theory.

Following the presentation a number of the points raised were discussed. In particular, the relation between belief ordering and the role of the preference and values system was debated. No firm conclusions on any of the questions raised were reached.

Models of Belief Revision

Aim: 7 issues relevant to alternatives

Initial assumptions:

1. Changes in cognitive state.
   Expansions + contractions of belief set.
   Non-monotonic reasoning.

2. A problem: Multiple extensions.
   - alternatives of belief revision, logically equivalent.

   eg. (a) $P \vee Q$ (b) $R \supset Q$ (c) $P \vee R$

   new evidence: $\sim P \wedge \sim Q$

   With new evidence - (b) or (c)

3. Obvious solution: Introduce ordering / priorities

2.

Issue 1 : Ordering / priorities of what?

Ordering beliefs - combinations as evidence for further belief

OR

Ordering rules of inference to infer further belief.

eg Individual beliefs with: probabilities,
support + plausibility ratings
endorsements,
degrees of confirmation .....

OR

Rules distinguished re specificity / generality -
preferring the most specific

eg. • mammals generally do not fly.
• bats generally do fly.
a is a bat. A bat is a mammal.

Issue 2 :

Ordering as predetermined and inherent in the structuring of the belief system

OR

Ordering as an end product of reasoning in the context of the belief system?

eg. Konolige - HAEL

OR

Gärdenfors - rationality postulates determining 'epistemic entrenchment'.

4.

Issue 3: The nature of belief ordering -

Variably certain beliefs

OR

Certain beliefs as variably corrigible,

[or entrenched
or hard to revise
or persistent.]

ie. beliefs with probabilities

OR

yes/no beliefs - variably useful in inquiry + deliberation eg. Chemical experiments

pragmatic - in context.

Harman - psychological plausibility

Harman + Doyle - utility

5.

Issue 4: The representation of ordering (rules or beliefs) - qualitative OR quantitative

Quantitative :- computationally-intractable?
precision?
Distinguishing uncertainty + ignorance?
Providing reasons - flexibility?
- Representational adequacy?

Qualitative :- Providing explicit representation of relevant factors to reason about uncertainty/certainty (Issue 2)
eg. endorsements (Cohen, Carver)
postulates (Gärdenfors)

Issue 5: The origins of the ordering:

Are there domain-independent principles - ordering related to properties

OR

is the semantics of the domain all important - ordering related to content?

eg. centrality of belief - numbers of relations (Quine) or core theories immune to change (Kuhn Lakatos)

Konolige - argumentation system.

Harman - principles of minimal change + maximal coherence.

Explanatory power + simplicity (Thagard Levi)

± subject matter

Source of evidence. 76 Specificity.

7.

Issue 6: The context in which ordering would operate. ie the alternative theories of rational belief change

foundation OR coherence

Foundation :. beliefs + justification

eg. RMS's. founded by self-evident beliefs (ass's), disbelief propagation

Coherence: mutually supporting beliefs

conservatism

minimal change + maximum coherence

Harman - positive undermining

8.

Issue 7: Modelling the context.

Formal models - computationally intractable for "real" problems

OR

Computational models - lacking semantic theory

Formal: eg. coherence models

Rao + Foo (possible words)

Nebel / Gärdenfors

coherence as logical consistency

Idealised rationality?

Computational: RMS's - tractable?

No interpretation

Logical inference reintroduced as justification

Emphasis on programmer - ad hoc?

Heuristics?

9

Issues - solving the multiple extensions problem in belief revision by ORDERING / assigning priorities :-

1. Ordering what? Beliefs or Rules.
2. Ordering how? Predetermined or Reasoned about.
3. Nature? Variably certain or Certain + variably corrigible.
4. Representation? qualitative or quantitative
5. Origin? domain specific or independent
6. Context? Foundational or Coherence theories of belief revision
7. Model? Formal or Computational emphasis.

Current research of Julia Galliers: Autonomous Belief Revision

Julia Galliers
University of Cambridge, Computer Lab.

The research issue is choice about changing belief. The research aim is to establish and model a principled theoretical basis by which rational agents autonomously choose whether and how to change their cognitive state. The context of primary interest is dialogue between cooperative yet autonomous participants, where neither is assumed to doggedly stick to current viewpoints, nor to abandon them when contradictions arise assuming the speaker to be reliably sincere and of greater knowledge. The purpose of the research is as a component of a model of dialogue in which utterances are planned according to a desire for a particular effect on another's belief state, but acknowledging the hearer's control over whether this effect be actually achieved. In addition it is assumed that neither participant be like an empty vessel waiting to be filled. So the model of dialogue is one of jointly negotiated belief *revisions*.

The issue of choice of changing belief is linked with aspects of strength or certainty. There are alternative approaches to this. Each individual belief may have associated probabilities or certainty values or support and plausibility ratings, which are numeric, having been derived via mathematical rules for combining evidence. Alternatively, non-numeric information such as whether an item is a result of default reasoning or not, or relatively specific or general, may be associated in some way with beliefs and/or rules of inference. Traditionally in AI, nonmonotonic reasoning systems consider beliefs as all equal for purposes of support and inference. The view adopted in this research has been that beliefs are held and represented equally, but comparative strengths in context can be reasoned about, if and when that context includes some challenge to an existing belief. The comparison is of relative persistence. Which would be the harder to revise, according to the adopted principles of revision and the specific context of the challenge? Such an approach considers certain beliefs as variably 'corrigible' in the context of other beliefs held at that time, as opposed to representing each belief as variably certain.

An ATMS has been used to generate alternative environments for reflection about potential revisions. The emphasis for comparison is on the combinations of assumptions which underlie reasons for a belief, as opposed to the supporting reasons themselves. According to Harman's principle of positive undermining:

- only stop believing a current belief if there are positive reasons to do so, and this does not include an absence of justification for that belief. Positive reasons are believing that all one's reasons for believing relied crucially on false assumptions.

*Research supported by a SERC IT fellowship.

The ATMS is a foundational mechanism which relies on the concept of justification. Foundation theory considers new beliefs are only to be added on the basis of other justified beliefs, and beliefs no longer justified are abandoned (contradicted by Harman above). Justification is not infinite however, and so there must also be beliefs justified in themselves, self-evident beliefs or assumptions, which are foundational in justifying others.

Harman's principle of positive undermining is not foundational but does express a view of belief sets in which some beliefs are related to others. Some beliefs are reasons for others; they are consistent beliefs related in being justifications or explanations of each other. They are coherent, mutually supporting beliefs which fit in well with everything else one believes. Existing formal models of coherence theory however, offer only logical consistency as the nature of this mutual support (exception : Gärdenfors' epistemic entrenchments).

The extensions generated by the ATMS comprise sets of consistent beliefs. Some of these are related by justification and some also by explanation because the examples considered include assumptions (and correspondingly beliefs inferred from them) which can be additionally justified themselves as potential explanations of other inferred beliefs. In fact, although the ATMS is a foundational mechanism relying on the concept of justification, the extensions are sets of *coherent* beliefs because in the sense of either being a justification, a possible explanation or else simply being consistent but not specially related, they are mutually supporting. I have attempted to adopt principles of coherence into this foundational model: minimal change for maximal gains in coherence. The harder belief to revise or the more persistent one, is the one residing in the preferred extension. This is the extension which would require more changes for less rewards in overall coherence to abandon, than any of the others.

The assessment of changes and coherence is on the basis of the sets of assumptions or alternative *interpretations* for each potential extension. What would be involved in believing all the reasons for a belief relied on false assumptions? Every assumption is represented along with an endorsement or indication of it's source. Overall coherence is determined according to a limited set of heuristics governing the combinations of these, such as that beliefs founded upon first-hand evidence are harder to disbelieve than those founded on any other combination of assumptions. (This doesn't take the possibility of faulty sensors into account). The theory is that there are general rules which can be brought into play and are relevant to choice, which are domain independent and related to the numbers and sources of combined assumptions (context) underlying a coherent set of beliefs.

The next stage of the research is to embed this model into a dialogue system designed for cooperative problem-solving. Each participant is envisaged as having detailed knowledge of different aspects of the problem, but needing information from the other to refine understanding of the overall context. Only together and by conveying information which confirms or alters previous inferences made on the basis of incomplete information, can an attainable problem description and it's potential solution be arrived at. It is suggested that knowledge of principles of autonomous belief revision drives the speaker's selection of appropriate intended belief states in conjunction with whatever knowledge is available about the hearer in particular, and of course the overall plan. If the participants in the dialogue understand the primary importance of assumptions grounding explanatory and justificatory reasons for beliefs, then the job of assisting the other to revise their beliefs is to find out or predict upon what assumptions their existing beliefs are based. Believing such assumptions false leads to dropping a belief, and the theory describes the basis upon which combinations of endorsed assumptions are dropped in favour of others. In addition,

assumptions can be suggested which would imply or explain (cohere better with) other data, and which would then lead to intended belief changes.

References

[1] Cohen P.R. Heuristic Reasoning about Uncertainty: an Artificial Intelligence Approach, Pitman, Boston, 1985.

[2] De Kleer J. An Assumption-based TMS. Artificial Intelligence Vol 28 No. 2 pp127-162,1986.

[3] Gärdenfors P. Knowledge in Flux: Modeling the Dynamics of Epistemic States. MIT Press, 1988.

[4] Harman G. Change in View - Principles in Reasoning. Bradford Book, MIT Press, Camb., Mass. 1986

[5] Levi I. Truth, Fallibility and the Growth of Knowledge. in Decisions and Revisions, Cambridge University Press, 1984.

[6] Rao A.S. and Foo N.Y. Minimal Change and Maximal Coherence: A Basis for Belief Revision and Reasoning about Actions. in Proceedings IJCAI '89, Detroit, U.S.A. 1989.

Current research of Mark Elsom-Cook: Educational agents in teaching systems

Mark Elsom-Cook
IET, Open University

The main focus of my research is on the development of computer- based tools for teaching. In particular, I wish to develop an interaction between computer and user which is essentially symmetrical. The computer tutor is an agent (which knows about its own internal structure) and the tutor models the learner on the assumption that it is a similar agent. This work is currently progressing in two directions (and it really is work in progress!):

1. Using an agent design for dialogue generation and understanding. The focus of this work is on the development of a mechanism for the generation of teaching actions which are justifiable by the system and motivated by a model of the learning process. The agent is provided with beliefs about a number of forms of conflict which can give rise to learning. From this knowledge and a desire for the student to know about a particular subject area, the system generates teaching strategies and dialogue actions consequent upon them. So far this work has involved implementing a simplified version of Kiss' agent architecture (with a trivial value system) and analysing human-human teaching interactions according to this model. The system has also successfully generated a teaching strategy, but no dialogue actions.

2. Using agents to build a model of the learner. Previous attempts at modelling a learner have essentially regarded the learner as a passive object to be filled with knowledge rather than an active constructor of theories. This work, which is still at a very early stage, is an attempt to combine Assumption-based Truth Maintenance Systems and machine learning models into a system which models the learner as an active problem-solver and theory builder. These components alone do not provide a motivation for the learner, nor a way of focussing on certain aspects of a problem and hence developing incomplete and inconsistent models. The intention is to use an agent architecture as the overall organising framework for this system.

SESSION 4:
THE VALUE OF FORMAL APPROACHES

PRESENTED BY: Nigel Shadbolt

REPORTED BY: Nigel Seel

Also reports of current work by:

Nigel Shadbolt on belief representation and agent architecture

Nigel Seel on Communication between agents

The Value of Formal Approaches

Presenter:Nigel Shadbolt

Rapporteur: Nigel Seel

In his introduction, Nigel Shadbolt expressed the problem as reconciling the approach of 'symbolic AI', with its emphasis on logic, formality and precision, with the psychological paradigm of 'artificial believers'.

He started by listing the benefits of the formal approach: a coherent theory of reasoning ("proof theory"), the separability of pragmatic aspects of reasoning from syntactic issues of deducibility (into a 'control theory'), the orthogonality of semantic concerns from both the preceeding (via notions of interpretation).

The formal approach has shown itself adaptable, by the use of non-standard logics for expressing intentional reasoning, and techniques such as abduction to capture non-deductive problem-solving operations.

Problems persist, however: particularly that of logical omniscience. Standard epistemic/doxastic logics commit the knower or believer to know or belief all the logical consequences of a set of basic beliefs. This contradicts our intuitions of the pragmatic, resource-bounded cognitive processing of implementable artificial believers. Three approaches have been tried:

1. the logical approach.

   Here one tries to amend the logical system to capture limited reasoning. Konolige introduced deduction structures, which might have incomplete inference rule sets, or resource-limited inferential systems; Levesque developed a logic of implicit and explicit belief, while Fagin and Halpern have worked on 'awareness logics'. None of these approaches has, however, been overwhelmingly compelling.

2. the procedural approach

   Ballim and Wilks (see eg Proc. IJCAI-87 pp 118-124 - "Multiple Agents and the Heuristic Ascription of Belief") present a program - ViewGen - which maintains a collection of embedded contexts, corresponding to the belief set of a particular believer. Thus "my beliefs are held in a box, my model of you as a believer is a box inside my box, in which your (assumed) beliefs are held". Such boxes may be nested arbitrarily. Ballim and Wils use data structures to model beliefs, including lambda expressions for beliefs which only other believers can evaluate ("I know you know John's telephone number, but I don't).

   This work is classical AI engineering, and it is an interesting question as to how easy it would be to reformnulate the work in a formal framework. This leads to the third possible approach.

3. the dual approach

This is work currently being investigated at Nottingham. The approach is to take a computational model, similar to the embedded contexts of ViewGen, but to use a formal language instead of computationally conveniant data structures to represent beliefs. The hope is that this mix of declarative and procedural approaches can begin to reconcile the logicist and psychological paradigms.

DISCUSSION

Steve Pullman (SRI) was concerned that the restricted constructs used in the Ballim and Wilks work generated insuperable problems. Thus if "For all x, Texan(x)" is represented by a 'generic Texan' data structure, then 'how tall is the generic Texan' ?

There was then some discussion about how the proposed 'dual approach' related to the earlier discussions about 'situated automata' vs 'symbolic AI'. The point was made that reflective thought seems to demand symbolic AI.

Finally, there was some discussion about reflexive belief, and the mechanisms by which such an artificial believer might model its own beliefs.

The value of formal approaches in the study of belief representation

Nigel Shadbolt
Artificial Intelligence Group,
Department of Psychology,
University of Nottingham,
Nottingham, NG7 2RD, UK
nrs@uk.ac.nott.psyc

Much of the work in cognitive science adopts what we might refer to as a formalist account of representation. A classic example is the analysis of meaning in language. The meaning of terms are construed as objective, model theoretic, entities set apart from cognitive considerations. How are we to relate such objects to psychology?

Many people candidly admit that the model theoretic frameworks entails that their work can have nothing to do with what does on in a persons head. We are studying mathematics not psychology. This disavowal of any interest in psychology and the cognitive system is one response that is open to scholars.

Other workers in the formal field have been extremely agitated by the gulf that separates them from cognition. They have placed on record their concern to bring together formal and psychological views of *semantics*. They assert that we will not be able to provide an adequate account of the propositional attitudes without a theory that reconciles logicist and psychologist.

In my view the fundamental issues of reconciliation have to do with whether the idealisations made by the logicists are compatible with a view of the propositional attitudes as psychological phenomena. My talk will examine these idealisations in some detail. It will attempt to find ways of making psychological sense of some and abandoning others for a different type of approach. I will describe some of the current work on developing representations of epistemic states and their processing underway at Nottingham.

The value of formal approaches to the study of belief representation

Nigel Shadbolt
Artificial Intelligence Group
Department of Psychology
University of Nottingham

22 March 1990
SRI Cambridge

Structure

- The power of logic
- What are the generic problems?
- Solutions to the generic worries?
- The particular problems of artificial believers?
- What to do?

The power of logic?

- A search for the language of thought
- The Physical Symbol Hypothesis
- Symbolic AI
- Logic as *the* KRL
  - language, semantics and proof theory (Reichgelt (In Press))
  - soundness and completeness
  - control
  - theory of meaning + expressivity

Problems and Responses

- Efficiency
  - Improved control
- FOPC Semi-decidable: loss of completeness
  - NB adopting heuristic control may loose soundness, for example default logic
- Imperative knowledge
  - Procedural
- Non-standard reasoning
  - Non-standard logics
  - Segregate the *extra-logical*, eg TEST

The particular problems of artificial believers?

- Epistemic Modal Logic
- Autoepistemic - reliance on negation as failure
- Superbelievers
- Logical omnisicience (Fagin and Halpern)
  - lack of awareness
  - resource boundedness
  - lack of inference rules
  - limited focus of attention

What to do?

- Logical approaches
- Procedural approaches
- A *Dual* approach

Logical approaches

- Konolige's syntactic approach
- Levesque - explicit and implicit belief
- Fagin and Halpern - Awareness logic

The procedural account

- Exemplar is Ballim and Wilks
- Use of viewpoints
- Radical opacity
- Default ascription heuristic
- Problems of; relevance, omniscience, conviction
- Lazy versus zealous/eager evaluation
- Issue of belief maintenance
- A pragmatic solution to logical omniscience

The procedural account within logic

- Can we reformulate Ballim & Wilks inside a logic?
- VIEWGEN as a mere implementation of such a logical description
- Thus Default Ascription Heuristic
  The predicate **true** maps the reified sentence **[x]** and a world index **w** to a truth value

$$\forall w \forall i \forall j \forall x((atom([x]) \wedge true(w, [B_i x] \wedge \neg contradict(w, [B_i B_j x])))$$

$$true(w, [B_i B_j x]))$$

where contradict is defined as

$$\forall w \forall i \forall x(true(w, [B_i \neg x]) \vee \exists y(true(w, [B_i y]) \wedge (B_i y \rightarrow$$

$$B_i \neg x)) \rightarrow contradicts(w, [B_i x]))$$

and note that atom excludes

$$\forall i \forall x(\neg atom((B_i x)))$$

- Loose the pragmatic solution to logical omniscience
- Reading of DAH is omniscient through the universal quantifiers

The dual approach

- The appeal of viewpoints
- Hardwiring prevents full range of modal epistemic logic
- How to represent?
  I believe all Texans think Texas is the best state in the Union
- Ballim & Wilks offer generic inheritance (instantiate 'a Texan' viewpoint and inherit
- Dual approach: Viewpoints + full logical language
- Two types of belief; declarative (implicit) and procedural (explicit)
- In some respects similar to Fagin and Halperns approach

The psychological irreducibility of *belief*

- Reasoning is localised
- Reasoning is goal directed
- Reasoning is reconstructive
- Reasoning is distributed
- Intensions are vague and negotiated
- Intensions are supported and rejected for different reasons
- Awareness is limited

Current research of Nigel Shadbolt: Belief Representation and Agent Architecture research at the University of Nottingham

Nigel Shadbolt
Psychology Dept., University of Nottingham

The work underway in the AI Group reflects the research interests of both myself and Han Reichgelt. A long standing interest of mine has been the the basis of communication between autonomous agents.

An ESRC grant (Project Number CO8250016 Cooperative planning: A foundation for communicative negotiation), awarded in 1985, looked at the interpretation and generation of cooperative dialogue between agents. The idea was to derive discourse from an underlying planning system (Shadbolt[4])- the thesis is that the fundamental driving force behind dialogue is the problem solving ability of the agent. The research has shown, via the construction of a number of computational models, that flexible dialogue can be obtained under this organisation (Shadbolt[5]). A number of important insights have arisen from the research. In particular, it became apparent that the planning system which was adopted for the computational modelling (NOAH Tate[6]) was too inflexible. It did not permit the sort of reflective reasoning an agent needs to carry out, reasoning which involves its own and other agents' states.

This led to the investigation of more powerful problem solving architectures (Reichgelt and Shadbolt[1], [2],[3],) which might provide the sort of reasoning outlined above, but which could also *compile* efficient problem solving solutions along the lines of traditional planning architectures. Much of this work has been carried out jointly with Han Reichgelt.

The three research projects that together provide the basis for our current work in this area are; SERC GR/F 28618 Epistemic Logic for Multi-Agent Planning Systems, SERC GR/F 35968 INFORMER - Integrated formalisms for epistemic reasoning, Joint Council Initiative in Cognitive Science and HCI SPG8826298 Planning and Instruction.

The first two of these are complementary projects The principal objective of the first is to understand the computational components required to produce autonomous knowledge-based sys-

tems capable of reflecting about their problem solving and *knowledge.* An important aspect of this problem is developing a system capable of reasoning about the model it maintains of the knowledge states of other systems.

The second project, which is due to start in April 1990 dovetails with the SERC project described above. It aims to provide efficient implementations of computational formalisms that allow knowledge-based systems to reason about the propositional attitudes.

The final project aims to integrate and develop empirical research in instruction with AI models of planning. The modelling of an instructional system capable of tracking the instructional state of a student provides a test bed for our interests in agent architectures, agent interaction and belief modelling.

References

[1] Reichgelt, H. and Shadbolt, N.R. (1989). Planning as Theory Extension. *Proceeding of AISB 89,* pp191-199, Pitman.

[2] Reichgelt, H. and Shadbolt, N.R. (1990). A Specification Tool for Planning Systems. *AI Group, University of Nottingham.*

[3] Reichgelt, H. and Shadbolt, N.R. (1990). TESTing Planning Systems. *AI Group, University of Nottingham.*

[4] Shadbolt, N.R. (1989). Planning and Discourse. In M.M. Taylor, F. Neel and D.G. Bouwhuis, Eds. *The Structure of Multimodal Dialogues.* North-Holland.

[5] Shadbolt, N.R. (1990). Speaking about Plans. *Proceedings of AICS'89.* Springer-Verlag.

[6] Tate, A. (1976) Project planning using a hierarchical non-linear planner. *Dept of AI Report 25. Edinburgh University.*

Current research of Nigel Seel: Communication between Agents

Nigel Seel
STC Technology Ltd., Essex CM17 9NA

Introduction

In recent work, ([See89], [See90a], [See90b]), I looked at a class of synchronous, object-oriented mathematical models capable of representing the interaction between an agent and its environment. I considered examples (in particular the case of an agent subject to psychological experiments in a Skinner Box) where an observer would be inclined to say that the agent was *intentional.* The agent would be considered to 'possess cognitive states' such as 'knowing things', and 'wanting things'.

Since the agent, like the objects in its environment, is just an automaton in the mathematical model, why is it being singled out for preferential treatment vis-a-vis its assumed intentionality ? I suggested that intentional descriptions[1] capture an important contingency in the way in which the agent is situated in its environment, in the following manner. There are things true in the world, which are of importance to the agent, but which (being contingent), were not able to be pre-programmed into the agent's design. If the agent can become aware (by perception) of such relevant facts, then it may be able to adjust its state so as to behave more appropriately.

In this view, intentional descriptions capture, in a non-architectural way, the state/process of attunement of a perceiving, acting and learning system to its environment. Note that it is the observer who is ascribing intentionality; the agent's operational mechanisms are just that, mechanisms. In the absence of an observer, there is no intentionality in the agent-environment setup.

When would *agents themselves* need to conceptualise other agents as being intentional ? Perhaps we should look to *social* agents, where the term 'social' is meant to capture coordination of behaviour ? Of course, to be a social agent is not to make very strong claims about an agent's cognitive prowess: both people and ants socially-coordinate behaviour, presumably with quite different cognitive apparatus.

Let's assume for discussion a collection of social agents, where the separate agents are capable of independently acting, and acquiring different information, which will have a bearing on the situation confronting the social group, and upon its success in determining and realising its objectives. So we may assume

1. a group-process of synthesis of multiple pieces of partial information derived from members of the social group,

2. a collective drawing of consequences as to the anticipated behaviour of the environment, and the possibilities for the group's future actions,

3. a group-decision about what is to be done, and a distribution of tasks amongst the group's members.

Suppose we take as the problem the simplest way to accomplish this. One way might be to describe the group in terms of a set of situation-assessment rules, something like:

[1] Formalised in an epistemic conative temporal logic.

If $observation_{j_1} \wedge ... \wedge$ $environmentState_j$ **thenNext** $environmentState_1$

... ...

If $observation_{j_1} \wedge ... \wedge$ $environmentState_j$ **thenNext** $environmentState_n$

and action rules:

If $environmentState_j \wedge groupSituation_j$ **thenNext** $groupaction_{ij}$

Given the limitations, partiality and perhaps error-proneness of perception, together with the possible limitations of such rule-sets in handling the complexity of the world, it would not be surprising if there was wide variance between the situation actually holding, and the group perception of the situation.

This 'epistemic gap', which leads to mistakes being made, generates the possibility of an epistemic-conative account of the group's cognitive situation just as for the solitary agent I discussed previously, and it seems likely that the technical approach I used then would still be applicable. I suspect that the interaction between group members would be 'ant-like', consisting of the exchange of tokens denoting observation and action classes, and would still not involve mutual recognition of intentionality. The changes in 'cognitive state' amongst the agents would presumably be restricted to selecting between a finite number of pre-programmed alternatives[2].

As an alternative, I think it is more promising to look at agent-agent interactions which are designed to alter the agents' cognitive states in a more open-ended way. Such interactions may be expected to be mediated by language which is expressive of both physical and cognitive states of affairs. The key concepts may be expected to include notions of conversation, language as action, negotiation, cooperation and conflict.

There is not much in place at the moment in terms of formalisations of these notions: I am currently looking at some of Barwise's analyses of common knowledge [Bar88], based on Aczel's work on non-well-founded set theory [Acz88], as well as Conversation Analysis and Speech-Act theory (see eg [Lev83]). I would like to derive an interaction model analogous to the one above for 'ant-like' social formations, show how the interactional requirements implied a need for mutual intentional modelling (ie appreciation that other agents should be treated as intentional), and then deduce what kind of 'language' is needed to enable such an interactional style, + what kind of agent architecture.

References

[Acz88] Non-Well-Founded Sets. P. Aczel. CSLI Lecture Notes No. 14. 1988.

[Bar88] The Situation in Logic - IV: On the Model Theory of Common Knowledge. J. Barwise. CSLI Report CSLI-88-122. 1988.

[Lev83] Pragmatics. S. C. Levinson. CUP. 1983.

[See89] "A Logic for Reactive System Design". Proc. AISB-89. N. R. Seel.

[See90a] "A Formalisation of First-Order Intentional Systems Theory". N. R. Seel, submitted to ECAI-90, 1990.

[See90b] Agent Theories and Architectures. Ph.D Thesis. Surrey University 1990. Also available as a technical report from STC Technology Ltd.

[2] I think it would be useful nevertheless to formalise an 'ant-like' social formation in a similar fashion to [See90a], if only to test these speculations.

SESSION 5:
KNOWLEDGE, ACTION, CHANCE AND UTILITY

PRESENTED BY: Sam Steel

REPORTED BY: Kave Eshgi

Also reports of current work by:

Sam Steel on Decision theory and modal logics of action and knowledge

Kave Eshgi on model-based diagnosis theory

Knowledge, Action, Chance and Utility

Presenter:Sam Steel

Rapporteur: Kave Eshgi

Sam Steel's talk focussed on the incorporation of decision theoretic notions into planning. He presented a framework in which the notion of expected utility is integrated with the semantic structures of Dynamic Action Logic. He also discussed how such a framework can be used for making rational plans.

One of the topics discussed following his talk was the relationship between his scheme and game-tree search methods using evaluation functions. Also the relationship with Rosenschein's Situated Automata theory was mentioned.

Decision theory & planning

Sam Steel

Dept Computer Science

University of Essex

"What shall I do?"

- action
- change (time)
- knowledge
  - knowing whether
  - [knowing how/what]
- [choice]
- utility
- chance

- planning to act
- planning to know
- [planning to plan]

Structure of talk

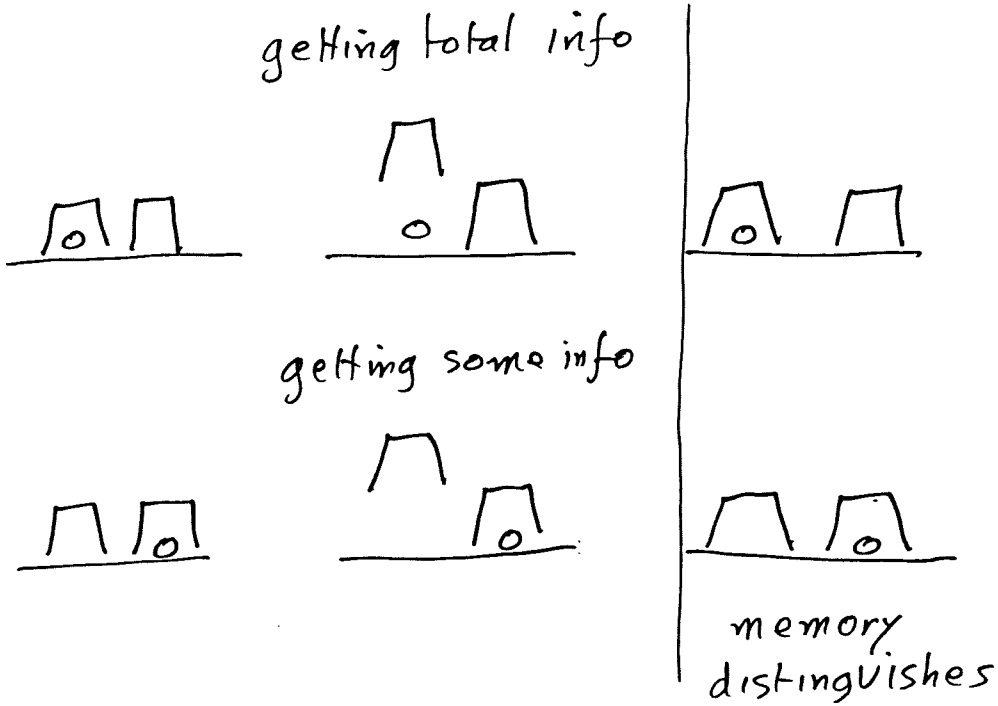
- Case: shell game
- Review decision theory
- adding D.T. to STRIPS
- review modal logic
  - knowledge
  - action
- adding D.T. to modal logic
- [planning based on that?]

Case

Knowledge / action

The shell game

Is there a pea? Where is it?



Obvious. Formalize!

Knowing how/what

Steel/Reichgelt

<no details>

not a good story to tell about

— why one must know how

— how one finds out how

Case

choice / action

1) fuel is gas
choose cooker
install chosen cooker

2) discover fuel (elec/gas)
choose cooker
install chosen cooker

3) choose cooker (elec)
choose cooker (gas)
discover fuel (elec/gas)
install chosen cooker

4) choose plan ((2)/(3))
execute chosen plan

5) C.B.P.

possible story available
<no details here>

Review

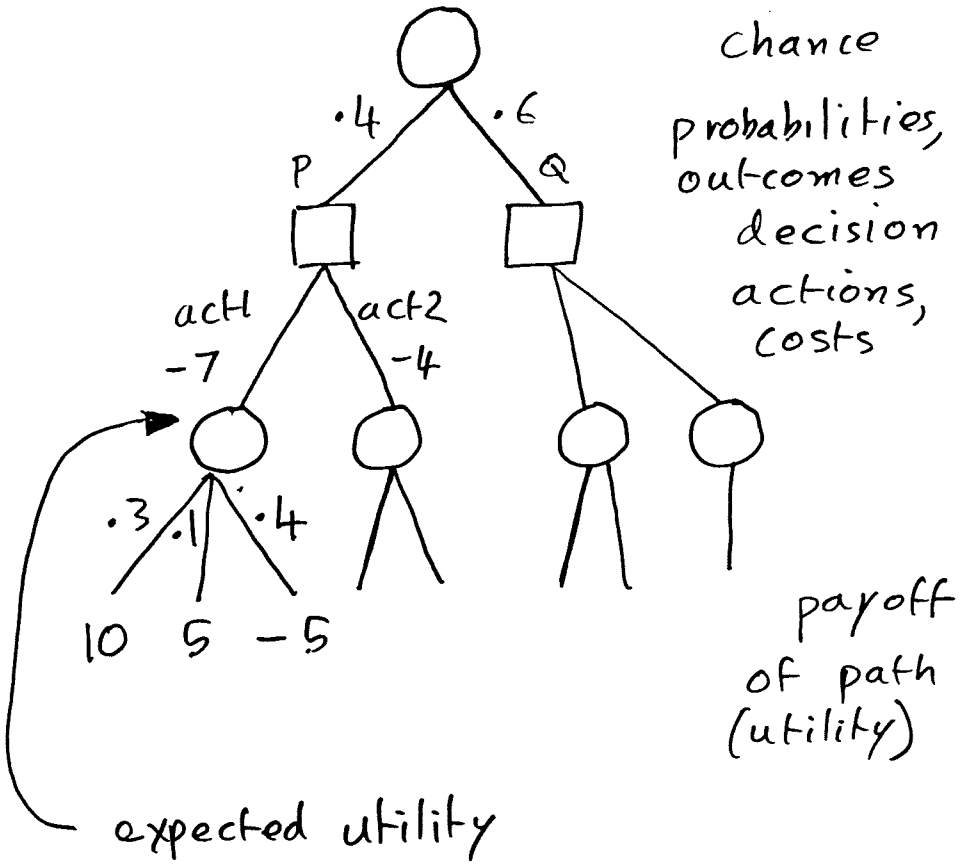
Decision theory

— special case of game theory

— extensive form

— strategies = programs = plans

extensive form

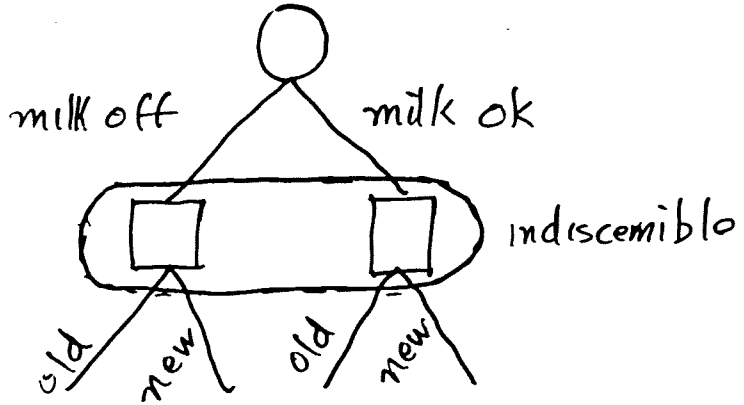


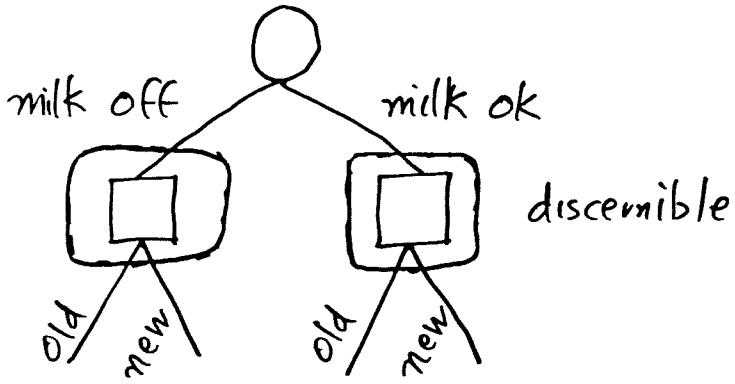
$.3*10 + .1*5 + .4*-5 = 1.5$

simple axioms imply

rationality = maximize expected utility

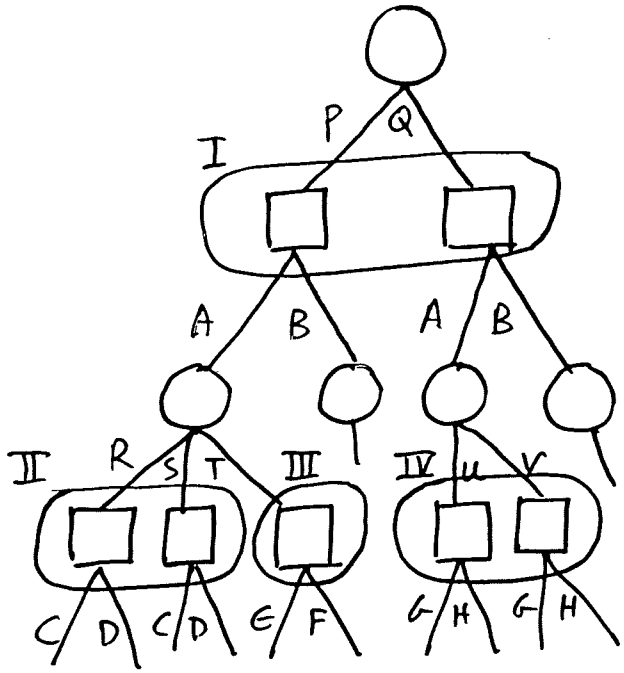
Information sets





Can only choose 1 action per info set

Strategy

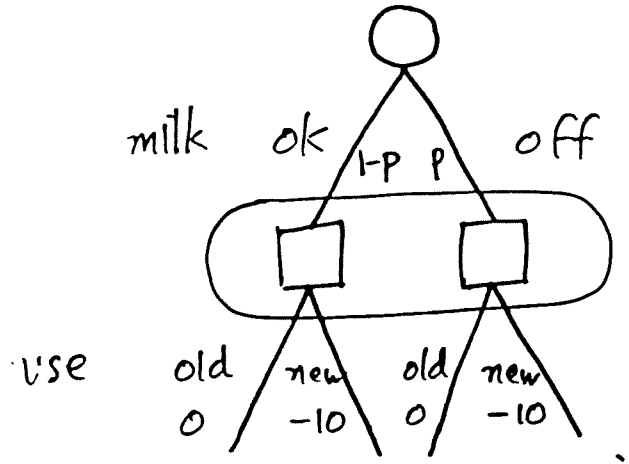


Strategy = rule about what to do at each info set
= function: InfoSets → Actions.

$\langle I \to A,$
$II \to C,$
$III \to E,$
$IV \to G,$
$\ldots\ldots$
$\rangle$

utility / probability / action

Is the milk off ?



utilities belong to strategies

use old: $-5*p + 10*(1-p) = 10-15p$

get new: $0*p + 0*(1-p) = 0$

use old if $0 < 10-15p \equiv 2/3 < p$

Section on decision theory & STRIPS omitted

STRIPS + DT:
good but not everything

— planning-to-know
vital but missing

— planning-to-know needs
sentences-as-arguments

— Believes(fred, $\forall x$(man(x)→mortal(x))

— one approach (others exist) is
K Sentence

— hence modal logic

Further
modal logic is nice for actions

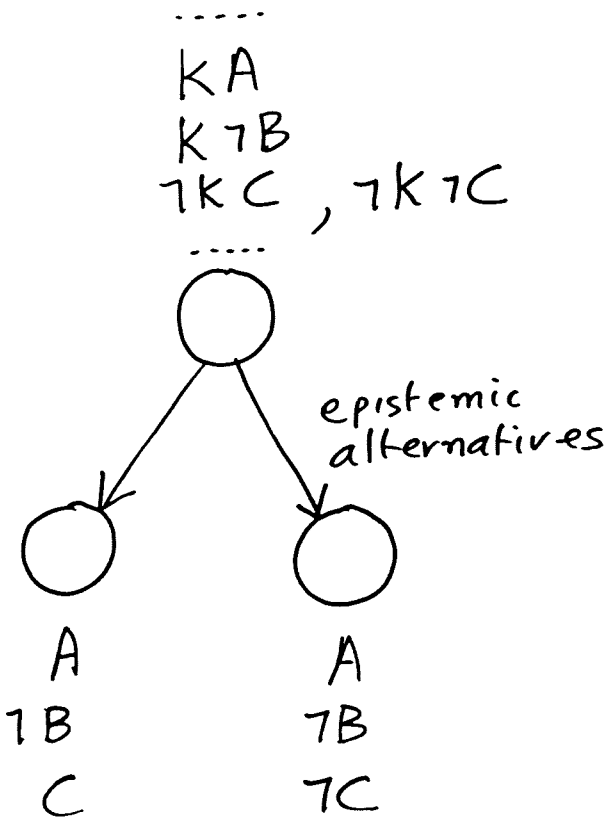
— concise (relative to sit. calc.)
(powerful but fiddly)

— no states; event-oriented
(common usage supports this)

— complex actions, eg
WHILE P DO A
are more easily expressed

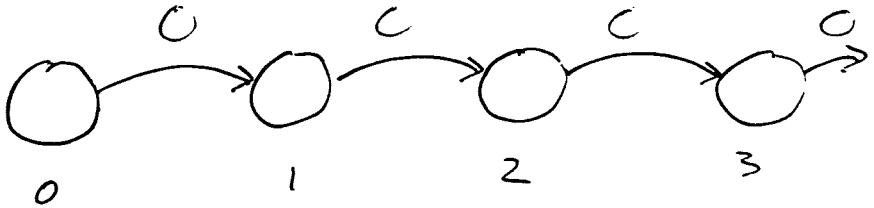
Review (epistemic) modal logic

K S in all credible worlds, S



Tense logic

world relation is / can be time step



- $[O]S$ — S is true next
- $[O^*]S$ — S is true hereafter
- $[(O^\cup)^*]S$ — S was true heretofore

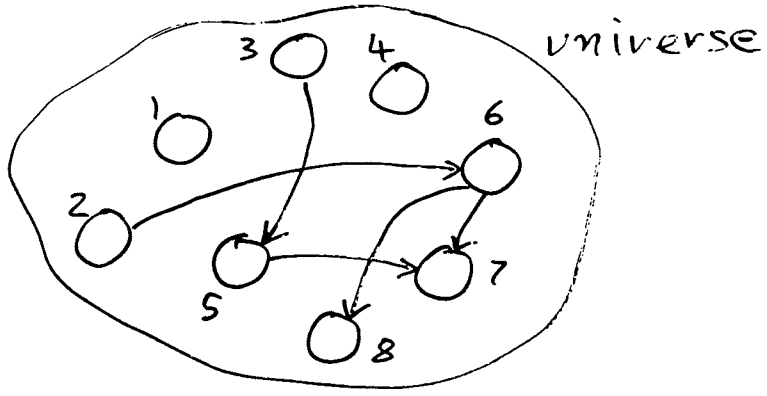
etc.

$FS =_{def} [O^*]S$

etc

Dynamic logic (Hoare → Pratt → Harel...)

Modal logic
with / accessibility relation / action type



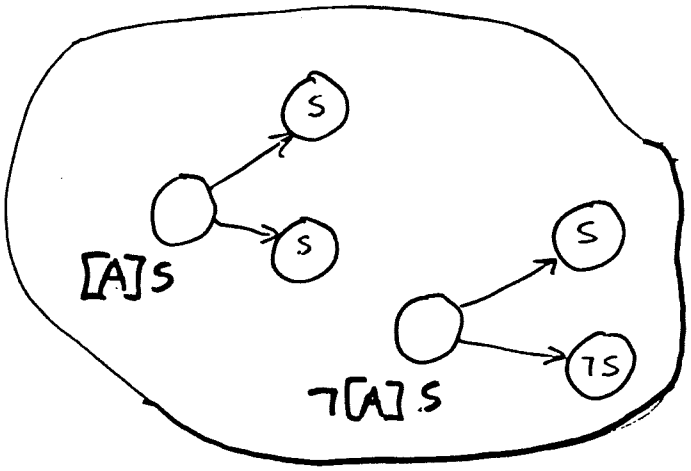
$[\![\text{Action}]\!] = \{ \langle 3,6 \rangle, \langle 3,5 \rangle, \ldots \}$

$[\![\text{Action}]\!]$ constant in universe
independent of world

[Act] S

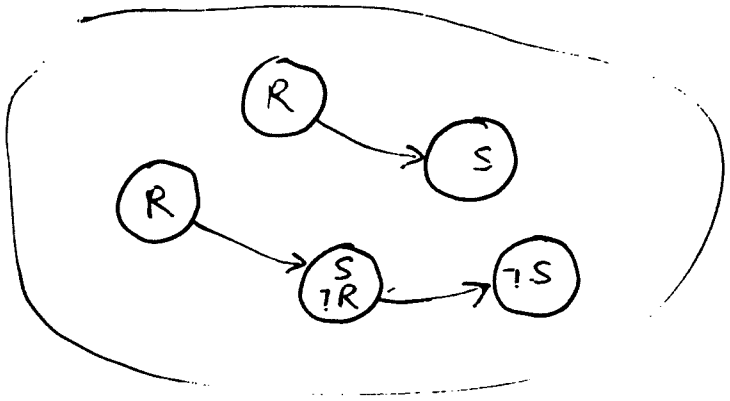
in all worlds that follow doing Act

S is true



$R \rightarrow [Act]\ S$

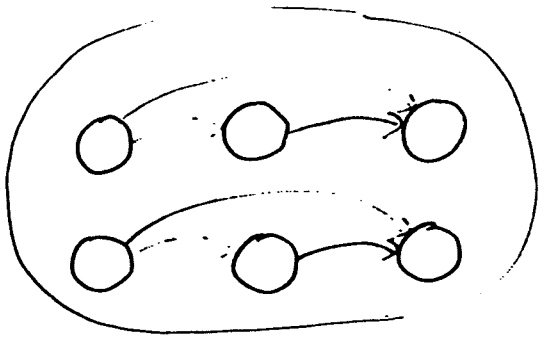
in all worlds where R is true
then in all worlds that follow doing Act
S is true



$R \rightarrow [Act]\ S$ is like operator

name: Act
preconds: R
effects : S

complex actions



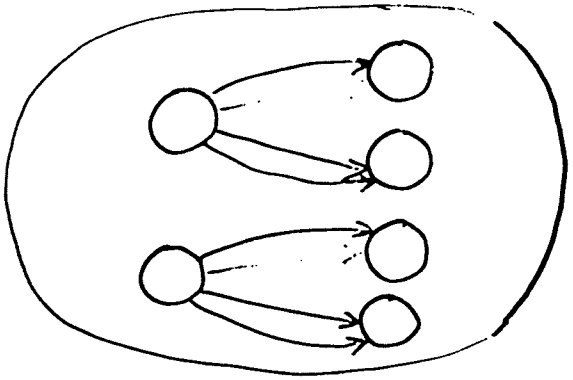
A

B

A;B

$[\![A;B]\!] = [\![A]\!] \circ [\![B]\!]$

$$\frac{[A]\,[B]\,S}{[A;B]\,S}$$



A

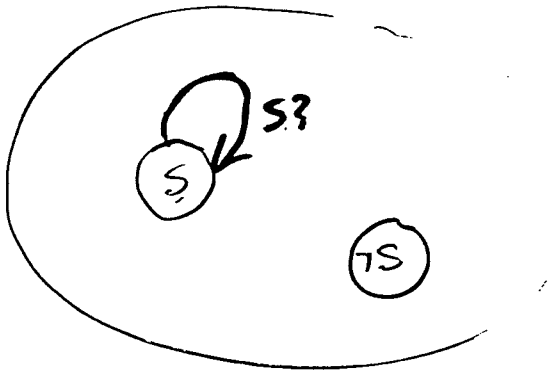
B

A ∪ B

$[\![A \cup B]\!] = [\![A]\!] \cup [\![B]\!]$

$$\frac{[A]\,S \quad [B]\,S}{[A \cup B]\,S}$$

a ~~test~~ guard $S?$



$$[\![S?]\!] = \{ \langle w, w \rangle \mid [\![S]\!] w = t \}$$

$$[S?]\,T \equiv S \to T$$

Non-termination

$[\![\text{Action}]\!] = \{\}$

No state follows doing Action

Termination

$[\![\text{Action}]\!] \neq \{\}$

$\text{Action}\downarrow$

$\text{Action}\downarrow \equiv \neg([\text{Action}]\#)$

Usually

non-termination =

running forever =

diverging

program constructs in dynamic logic

Do A only if P

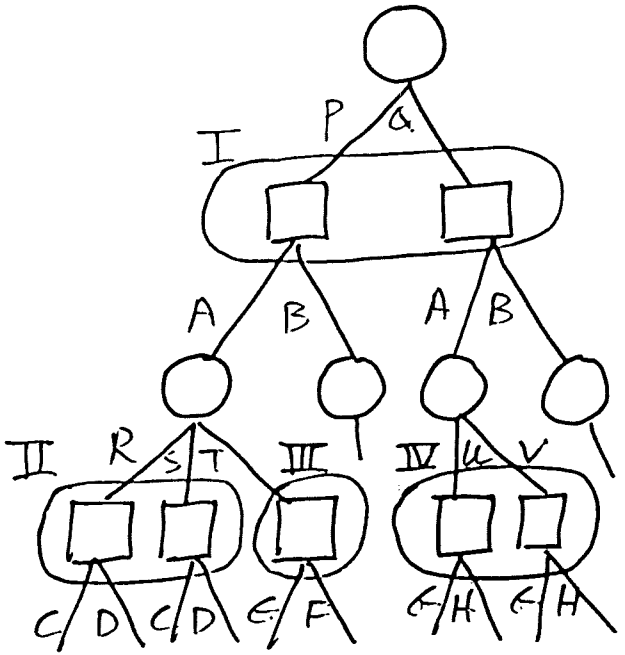
$P?;A$

If P then A else B

$(P?;A) \cup (\neg P?;B)$

While P do A

$\doteq (P?;A)^*;\neg P?$



strategy as function	strategy as program

$\langle I \to A,\ II \to C,\ III \to E,\ IV \to G,\ \ldots \rangle$

$(P \vee Q)?;\ A;\ ((R \vee S)?;\ C;\ \ldots)\ \cup\ (T?;\ E;\ \ldots)\ \cup\ ((U \vee V)?;\ G;\ \ldots)$

now

strategy = program

but (so far)

- no probability
- no utility

probability – Kolmogorov

Ω – all possible outcomes

$A, B, \ldots$ – subsets of outcomes

$0 \leq P(A)$

$P(\Omega) = 1$

A, B disjoint implies

$P(A) + P(B) = P(A \cup B)$

and for countably many $A, B, \ldots$

Measure space.

$\langle \Omega, S, m \rangle$

Ω — set

$S \subseteq \mathcal{P}\Omega$

Boolean algebra $(\cap, \cup, \setminus)$
closed under countable unions

$m : S \to \mathbb{R}$

measure of size of set
positive
countably additive

$m(\bigcup_i A_i) = \sum_i m(A_i)$

probability space: $m(\Omega) = 1$

$$P(A) = \left[\frac{m(A)}{m(\Omega)} = \frac{m(A)}{1} \right] = m(A)$$

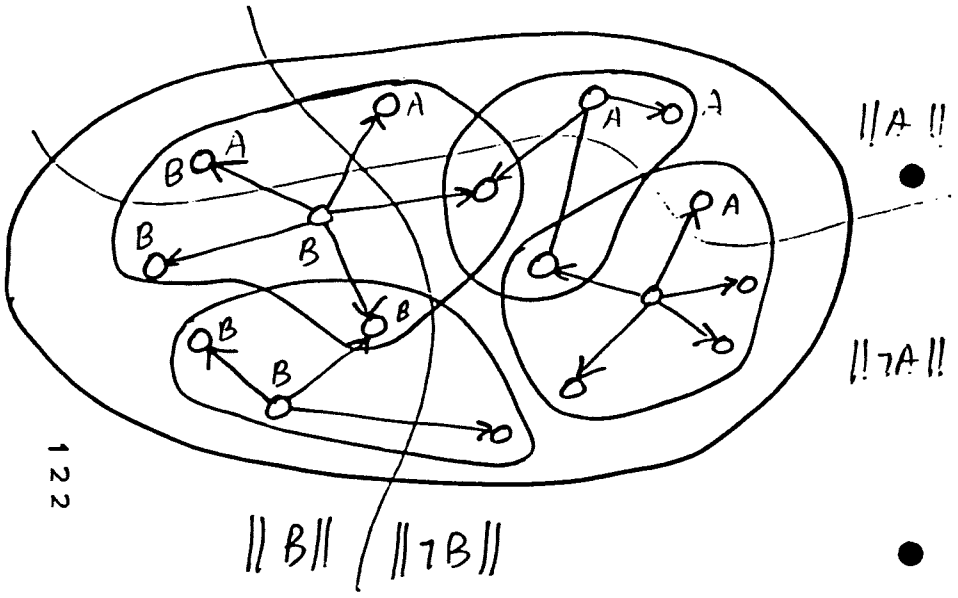
Adding probability to epistemic log.

Carnap's intuition

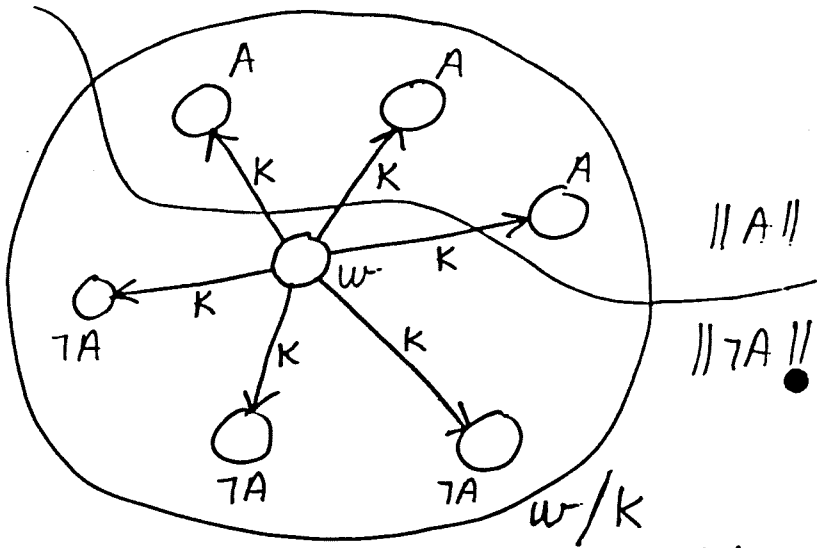
$P(A) =$

$$\frac{\text{number of worlds where } A \text{ is true}}{\text{number of worlds possible}}$$

Kripke structure on worlds +
measure space on worlds
=
model for probabilistic epistemic logic



Close-up



credible worlds around w

$$[\![P(A)]\!] w = \frac{m(\|A\| \cap w/k)}{m(\quad\quad w/k)}$$

$$[\![P(A|B)]\!] w = \frac{m(\|A\| \cap \|B\| \cap w/k)}{m(\quad\quad \|B\| \cap w/k)}$$

What follows?

"Kolmogorov axioms"

$\vDash P(\top) = 1$

$\vDash 0 \leq P(A)$

$K \neg(A \& B) \vDash P(A) + P(B) = P(A \vee B)$

Also

$K(A \equiv B) \vDash P(A) = P(B)$

$\vDash P(A|B) * P(B) = P(B|A) * P(A)$

$K(B) \vDash P(A|B) = P(A)$

$K(A \to B) \vDash P(A) \leq P(B)$

$\vDash P(A \to B) = P(\neg A) + P(B|A) * P(A)$

etc

higher-order probabilities clear

Expectations

X is a random variable $=_{def}$

$X : (\text{world } \Omega) \to \mathbb{R}$

$E(X) =$

average of X over possibilities $=$

(usually)

$\Sigma \; X(w) . P(\{w\}) \quad =$

$\Sigma \; X(w) . m(\{w\})$

in this formulation

$$[\![E(X)]\!]\, w =$$

$$\frac{\Sigma w' (w' \in w/K,\ [\![X]\!] w' \cdot m(\{w'\}))}{\Sigma w' (w' \in w/K,\ m(\{w\}))}$$

$$=$$

$$\frac{\Sigma w' (w' \in w/K,\ [\![X]\!] w' \cdot m(\{w'\}))}{\Sigma w' (w \in w/K,\ m(\{w\}))}$$

Also conditional expectation

$$E(X|A)$$

Main rule

$$K \vdash \neg(A \,\&\, B) \models$$

$$E(X|A \vee B) = E(X|A) * P(A) + E(X|B) * P(B)$$

In (countable) generality

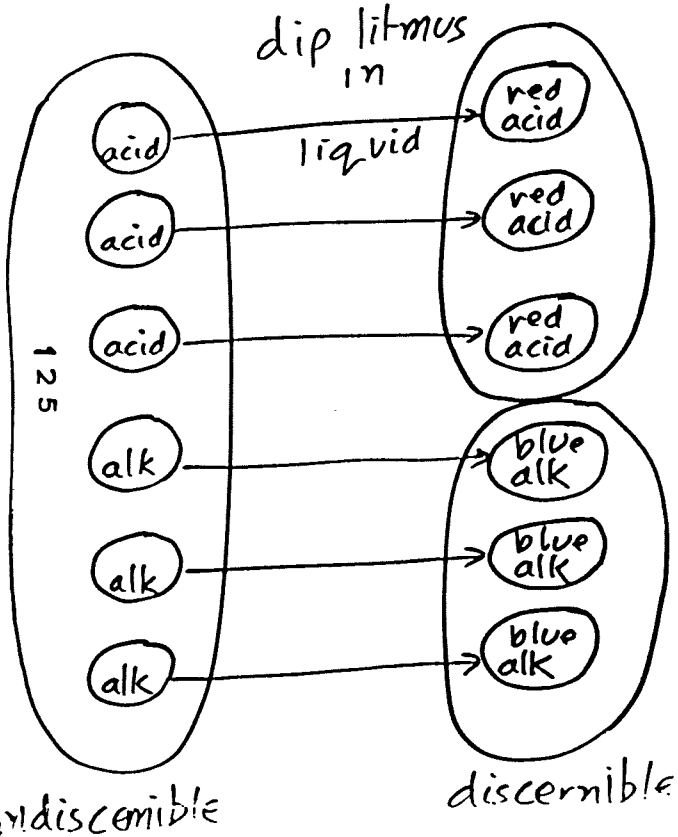
A_i mutually exclusive $\models$

$$E(X|\bigvee_i A_i) = \Sigma_i \left(E(X|A_i) * P(A_i) \right)$$

Interaction of action / knowledge

Moore's intuition

unknown liquid + litmus

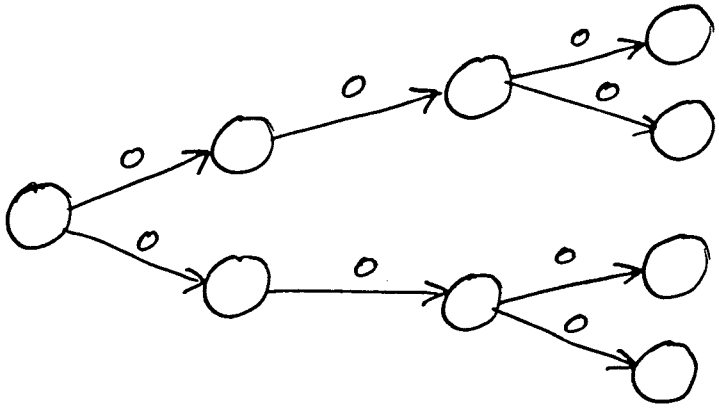


we know where we are; so we know where we started

Alternative (wrong) view

uncertain future → branching time

uncertainty real, not informational



Path modals eg

– on all paths, henceforth ....

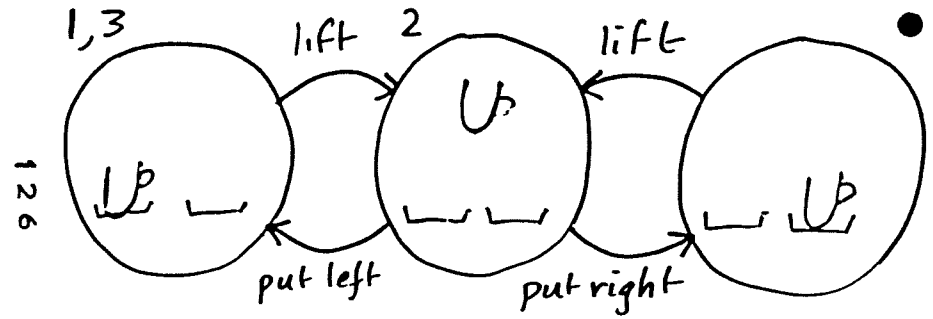
– on some path, eventually ....

etc

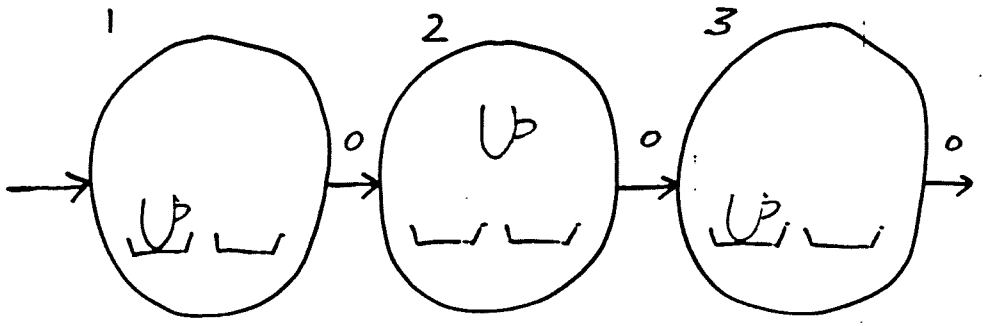
(some useful applications)

change of state ≠ change of time
both needed. consider
lifting then replacing tea cup

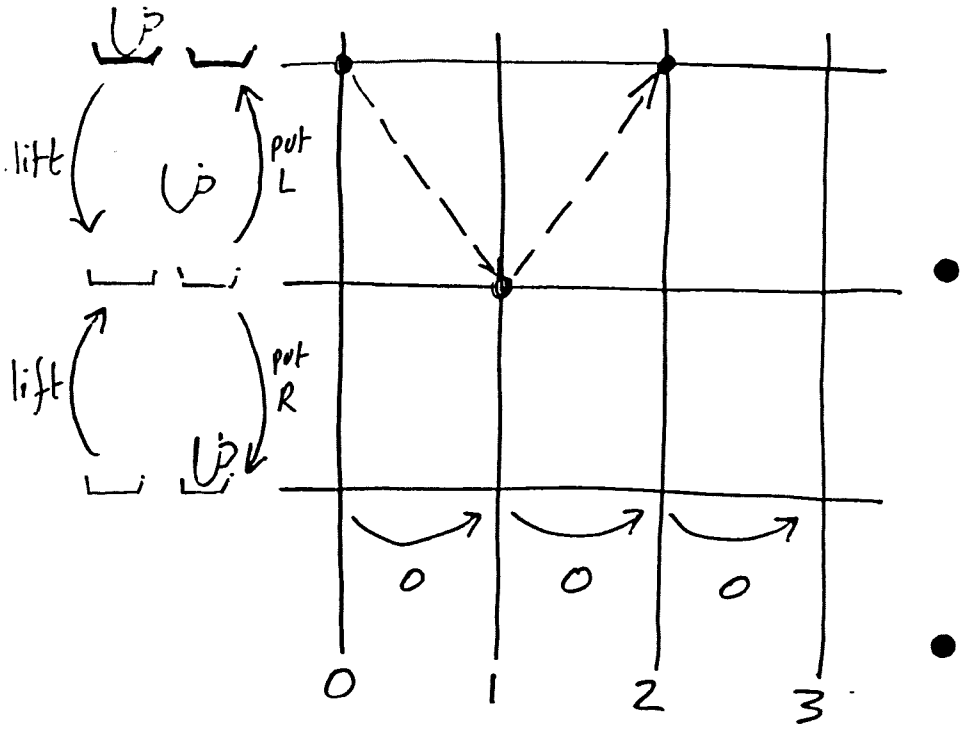
— state models
preclude memory of change



— time models
preclude sameness of state

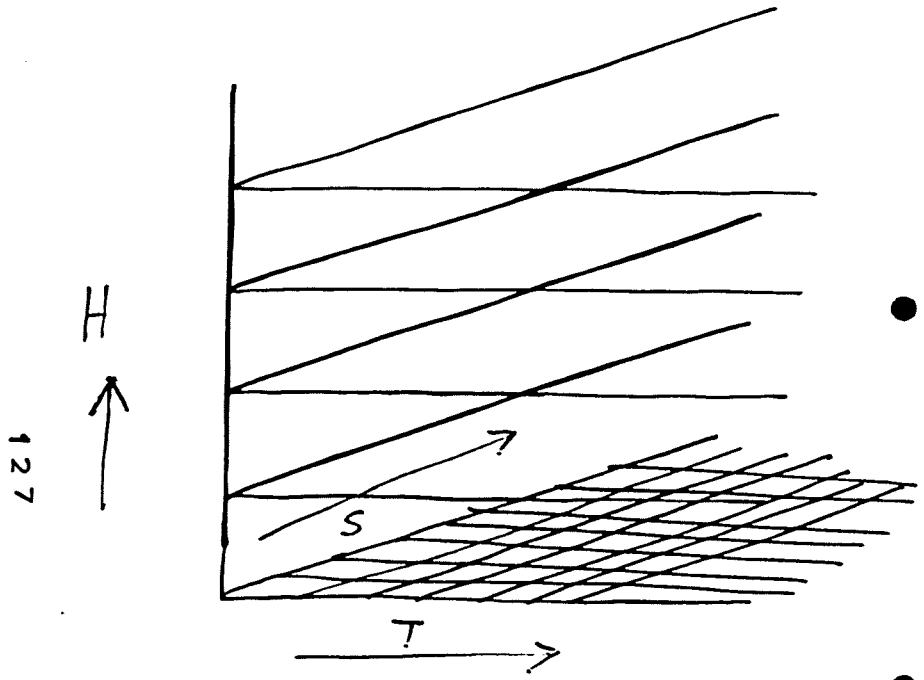


Idea: "phase space"
product of states and times



states are S X T
primitive actions are
(actions on S) X (intervals on T)
(lift X O)

alternative trajectories in S×T
lie on separate "sheets" = histories

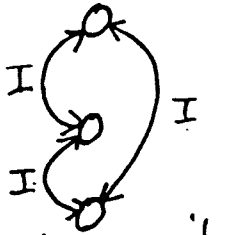


useful space is S×T×H

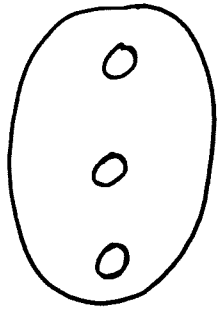
(Moore constructs histories,
rather than representing them)

uncertainty --
being somewhere in a volume
of points in S×T×H

Alternative reprs. of uncertainty sets



I = indiscernible



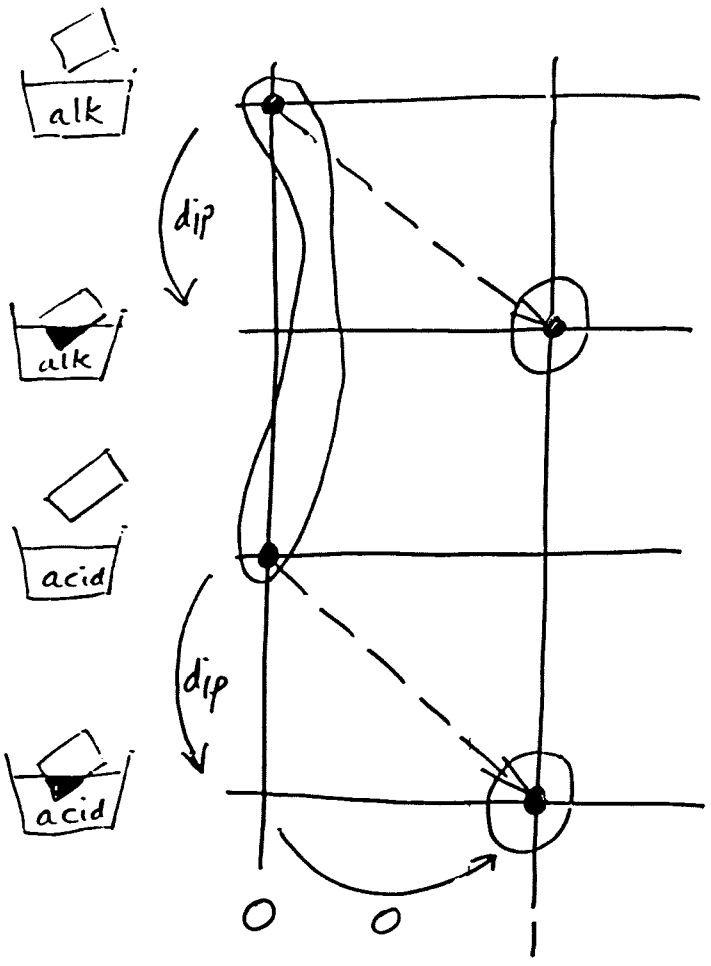
if I is equivalence relation
leads to S5

(other constraints on I possible)

I = K for epistemic models

Litmus example

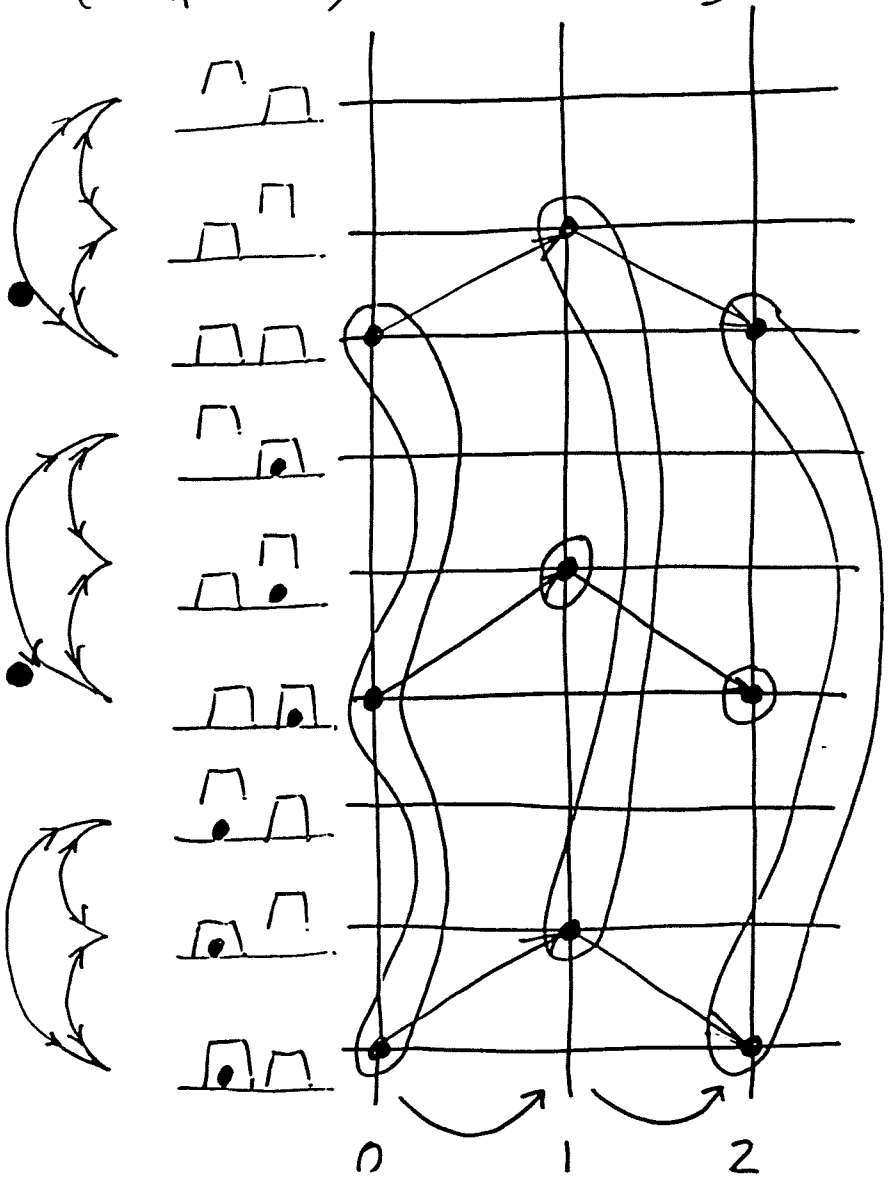
2 histories coalesced



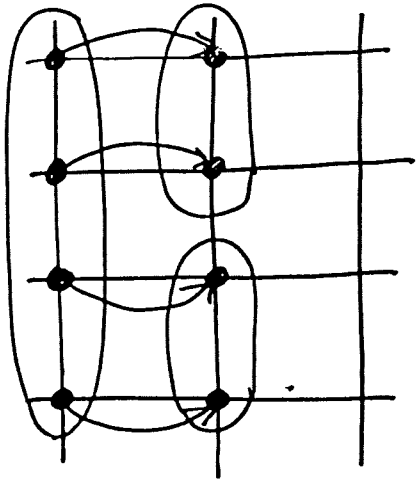
dip × 0

Memory in the shell game

Distinguished histories may stay distinct
(stipulated, not intrinsic)

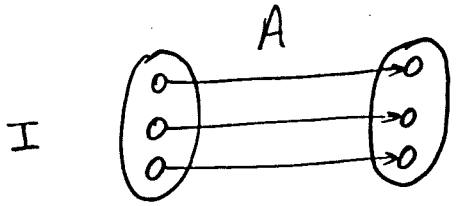


getting news
- info change
- no state change



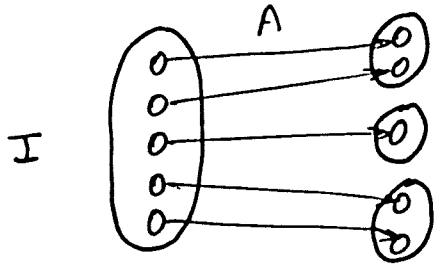
(≡ choice of a set member
≡ assuming a fact)

info. change



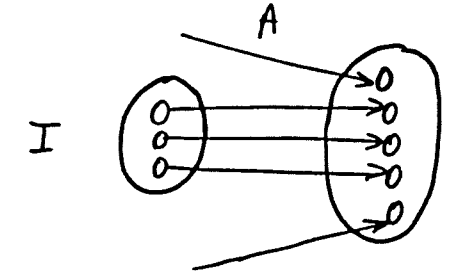
constant info.

$I \circ A = A \circ I$



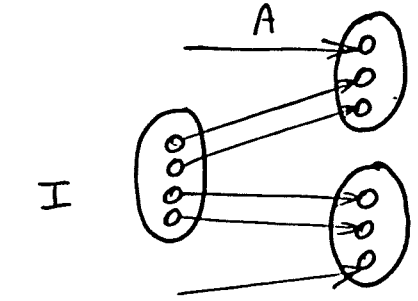
info. gain
(non-loss)

$I \circ A \subseteq A \circ I$



info. loss
(non-gain)

$I \circ A \supseteq A \circ I$



info revision (?)

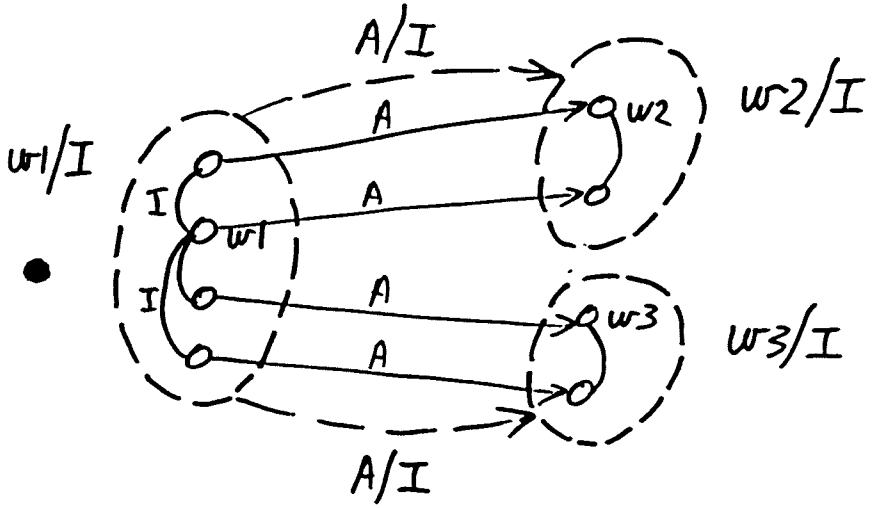
neither

(entropy quantifies info change)

- All actions are functions
- impossibility failures
  - block screwed down
  - change of time
  - no change of state;
  - = non-op = skip
- chaotic failures
  - dropping tea tray
  - uncertain outcomes due to uncertain start state

All randomness is epistemic
informational
mental

Quotient structures



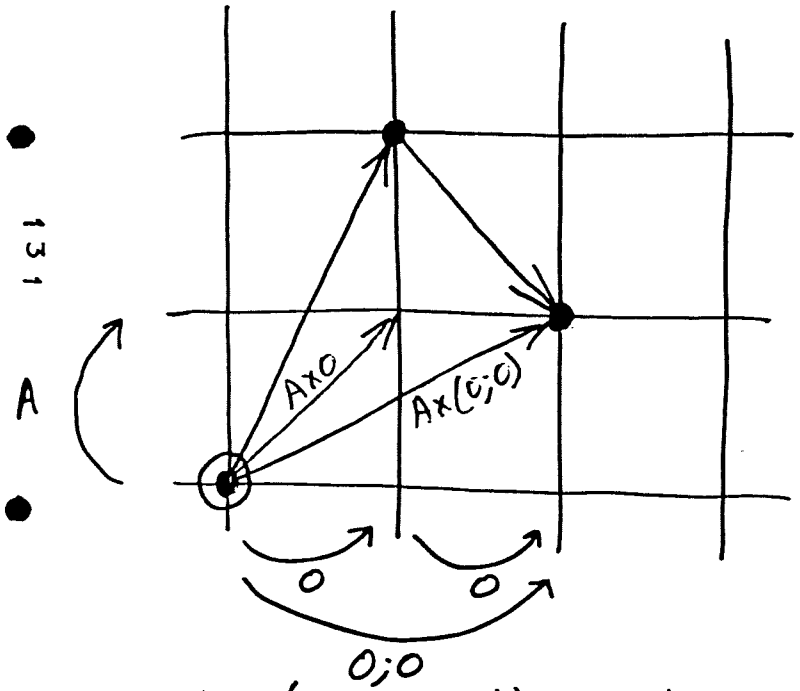
$(S \times T \times H)/I$ = "info. sets"
= dynamic logic states

A/I = relation on "info sets"
= dynamic logic actions

(& hence kozen?)

occ A

true where A begins

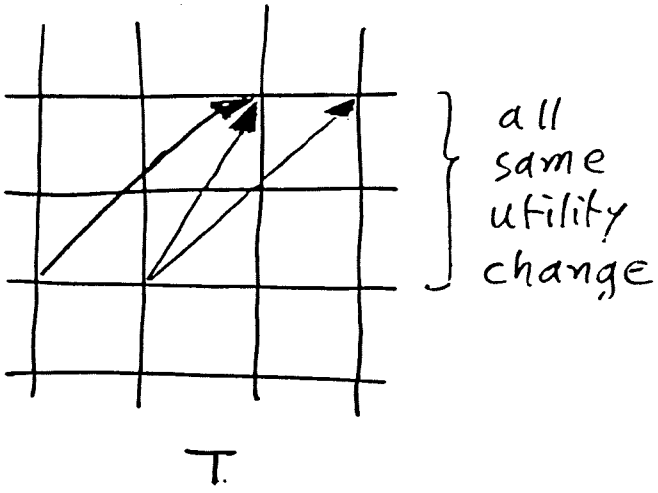


occ (A × (0;0)) true

occ (A × 0) false

occ X ≡def ¬[X]#

Utility



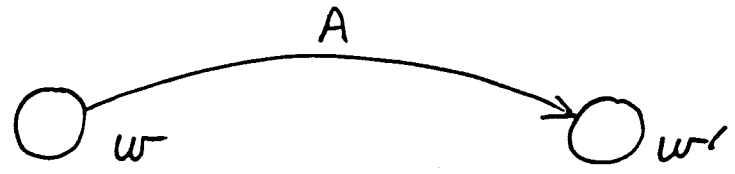
- utility in S×T×H =
- utility of s component

$g(\langle s,t,h\rangle) = g'(s)$

utility of action A at w

utility of A =

utility (post A) − utility (pre A)



$[\![\$A]\!] w = g(w') - g(w)$

$= g([\![A]\!](w)) - g(w)$

only defined if $[\![A]\!]$ is functional

expected utility of action A =

$E(\$A)$

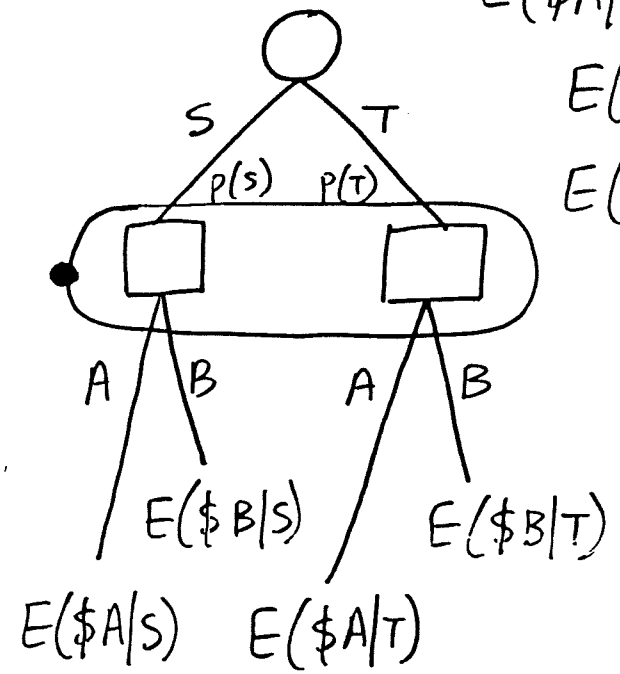
— — — dependent on circs.

$E(\$A \mid P)$

$E(\$A \mid S \vee T) =$

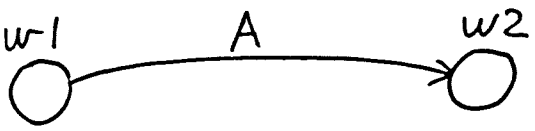
$E(\$A \mid S) * P(S) +$

$E(\$A \mid T) * P(T)$



provable

if all terms defined ....
all actions functional
all actions preserve measure

ie 

$$m(\{w1\}) = m(\{[\![A]\!](w1)\})$$

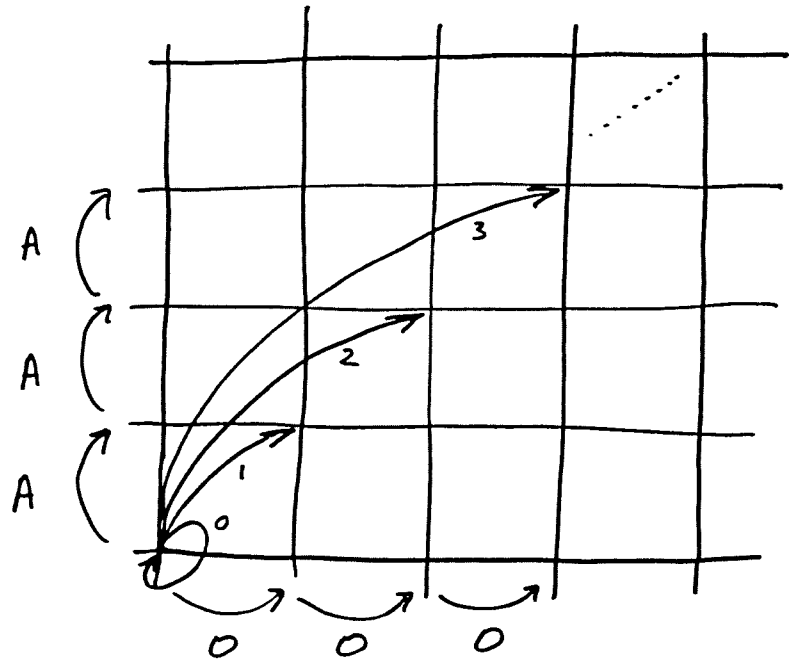
provable

- $\models \$\in? = 0$
- $A \circ I \subseteq I \circ A$ (info not lost),
  $\$A = x$
  $[A]\$B = y$
  $\models \$(A;B) = x+y$
- $\neg(occ(A \cap B))$ (A,B cant co-occur)
  $\models (\$(A \cup B) = \$A) \vee (\$(A \cup B) = \$B)$

A^* undefined

because A^* not functional

$occ\ A^* \models occ\ A^{n+1} \models occ\ A^n \ldots$



$(A \times O)^*$

"recursive games" have same problem

? axiomatize utility of action?

expected utilities of actions

- $E(\$\in ?) = 0$
- $E(\$A) = x$,
  $[A] \in (\$B) = y$,
  $E(\$A) = E(\$A \mid occ\ (A;B))$,
  ....
  $\models E(\$(A;B)) = x+y$
- $E(\$\ A \cup B) =$
  $E(\$A) * p(occ\ A) + \in(\$B) * p(occ\ B)$

leads into randomized strategies

so

utility/probability facts of decision theory can be replicated in modal logic

but

- plan construction ???
  (Rosenschein/Kautz)
- choice between
  planning/execution ???
  (Russell-Wefald)

Recent ideas (20/2/90)

Let utility be a "real" vector
= commodity bundle
≠ $\mathbb{R}$

Then

- costs of actions, merits of end states } coalesce
- utilities can reflect metrics
  tiredness
  niceness of tea
- Utilities superimpose thus

$$\$_{tired}\, A + \$_{nice}\, A = \$_{tired+nice}\, A$$

$$\sum_i \$_{m_i}\, A = \$_{\sum_i m_i}\, A$$

Utility of A^* ~~~ Markov chains

link ?

Because

$$A^* = \bigcup_{0 \le n} A^n$$

then show

i.e. $\$A^* = \sum_{0 \le n} \$A^n . p(occ\ A^n)$

If

$$p(occ\ A^{n+1} \mid occ^n\ A) = x$$

then

$$\$A^* = \sum_{0 \le n} \$A^n . x^n . (1-x)$$

~~Dynamic logic~~ ?

"Denotational semantics of programs"

construct

dynamic logic

D.S. of P.

A; B

A; B

A; B

if T then A else B

$(T?;A) \cup (\neg T?;B)$ abolish $\cup$!

$\{\langle t, A\rangle, \langle f, B\rangle\}(c)$ cases

while T do A

$(T?;A)^*;\neg T$

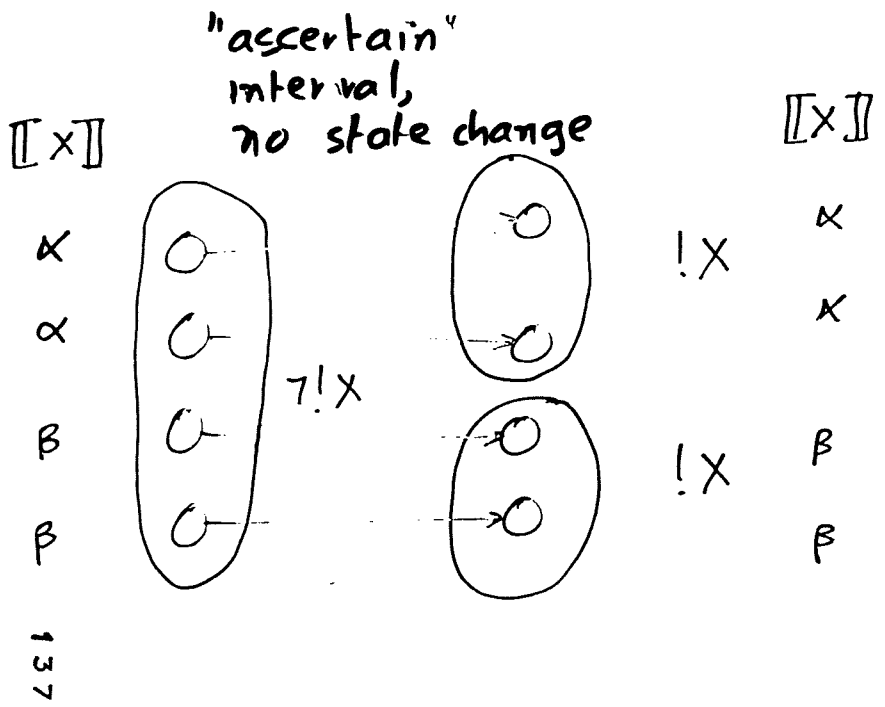
fix $(\lambda\theta.$ if T then A; θ else skip$)$

A || B

$A \cap B$????

?

The nature of choice



ascertain

— mary's phone number

[asc. phonenum(mary)] ! phonenum(mary)

may of course not terminate

— an action for a goal G

ε is a choice function, ie

Set ≠ {} ↔ (εS) ∈ S

[asc. ε(λA.[A]G)] ! ε(λA.[A]G)

[————————][ε(λA.[A]G] A

— best action
action of highest utility
(if one can cure log. omniscience)

[choose max(λA. $A)] ! max(λA. $A)

Current research of Sam Steel: Decision Theory and Modal Logics of Knowledge and Action

Sam Steel
Computer Science Dept., University of Essex

Abstract:

Decision theory talks about the interaction of knowledge and action, about probability and probable effects and utility of action. AI planning talks about the preconditions and effects of actions, the relation of atomic and complex actions and the semantics of change.

Those are different, but there is a large overlap between them. Putting them together would be good as a way of seeing how they relate, and as a way of putting measures of utility into standard AI planning. I shall suggest a way of doing this, by combining epistemic and dynamic modal logic, complicating the models so that they can also represent probability and utility, and then translating decision trees into that logic.

These ideas are still being worked out. I hope people will criticize and improve them.

This is an outline grant proposal that I and Han Reichgelt have made to the joint SERC/ESRC/MRC cognitive science initiative.

Computer models of knowledge and choice in planning

Summary of proposal.

To understand more about rational action, especially how it depends on knowledge, probability and choice. This will be done by writing programs to make plans that combine ideas from decision theory, logics of knowledge and action, and AI planning.

Here is a problem in rational action. A baby is about to be born. It will need a name. There are two strategies.

* Wait till its sex is known, then choose an appropriate name.

* Choose a name for each sex, and apply the right one when the sex is known.

These plans involve interleaving action on the world with choices about the operands of those actions. Furthermore, one has to choose between those two strategies. There is currently no well-understood way of making or even representing such plans. (This is the problem we address). We believe that the elements needed are in fact available, but in separated areas.

* Modal logic provides a logically impeccable medium for representation of knowledge and action, in epistemic logic [Hintikka] and dynamic logic [Harel]. Since those have common models, they can be combined in one language.

* Plans for action can be built using dynamic logic as the representation, either as in [Rosenschein] or by replicating STRIPS operators and doing non-linear planning [Tate, Chapman].

* Standard decision theory (eg [Raiffa]), based on subjective probability and utility, can model the perceived benefits of actions under uncertainty and preference.

So far is standard. But

* One can impose a probability space on models of modal logic [Halpern, Steel] and so let probability and mathematical expectations (eg of the utility of actions) into the logic. Then decision theory is available in modal logic, and strategies in the sense of decision theory can be identified with programs in DL.

* Using a particular sort of model for epistemic logics, we believe it is possible to represent the action of choosing. Essentially it appears as dividing a set of epistemically indistinguishable worlds into subsets. Each subset corresponds to having chosen one of the possible alternatives.

* The same framework supports a novel account of "knowing how" [Steel & Reichgelt] and of how gain in information justifies confidence in a plan. Essentially one "knows how" to do an action A if the denotation of A at any world one might be in will achieve one's goal whatever world one is actually in. This approach is compatible with [Moore, Morgenstern] but offers as theorems what they take as axioms.

* Recent work [Stuart & Wefald] has suggested ways of comparing the utilities of executing the action that currently seems best, and of estimating the expected utility of other actions too. Such estimation has a cost which must be weighed against any improvement found. We believe that this can be replicated in the framework we propose - it appears as attaching a cost to a choice among a set (of uncertain extension) of "optimal actions". This is how the choice among the two example in the baby-naming example is to be made.

This very compressed account is intended to suggest that the representation of subtle aspects of rational action in modal logic is possible. It is possible to construct plans in the related dynamic logic. There are two things that need to be done. (These are the contribution of the proposed work.)

* Showing that the representation is in fact adequate to describe a range of complex but realistic plans involving knowledge, probability and choice;

* Showing that such plans can be not just represented, but also constructed automatically. Post-hoc analysis of examples is important but leaves unanswered the objection that the choices that actually have to be made during planning are quite different and are made differently.

In order to show this we ask for a research assistant for 2 years and a Sun workstation (to be based at Essex) and for travel money between Essex and Nottingham. We will then write experimental planning programs using the ideas described. A very bald programme runs:

- implement a domain-action-only dynamic logic planner;

- add probability and utility measures to states and actions, to replicate the building of strategies from simple decision theory;

- separately add knowledge modals to the base system to produce a planner that will seek knowledge in order to "know how";

- combine the last two and add a choice operation so that the two different baby-naming plans can be created;

- incorporate the cost of choice so that the choice between those two plans can be made rationally.

Applications of this work are speculative. However applied systems, eg [Georgeff & Ingrand] are increasingly concerned with deciding whether to plan on current information or to spend effort on getting more information. If one can list in advance those facts which may be in doubt and which matter, then one can build in the proper choices. But a sufficiently autonomous agent, even a spacecraft worrying about the relative merits of repairing damaged sensors and dead-reckoning, must have a general rational process for making such choices.

REFERENCES

Chapman D: 1985
Planning for conjunctive goals
AI 32 (1987) 333:377

Georgeff, M; Ingrand, FF: 1989
Decision making in an embedded reasoning system
IJCAI-89 972-978

Halpern, JY: 1989
An analysis of first-order logics of probability
IJCAI-89 1375-1381

Harel, D: 1979
First-order dynamic logic
LNCS 68

Hintikka, J: 1962
Knowledge and belief
Cornell UP: Ithaca, NY

Moore, Robert: 1985
A formal theory of knowledge and action
CSLI-85-31, CSLI, Stanford U

Morgenstern, Leora: 1987
Knowledge preconditions for actions and plans
IJCAI-87 867-874

Raiffa, Howard: 1968
Decision analysis
Addison Wesley

Rosenschein, Stanley J: 1981
Plan synthesis: a logical perspective
IJCAI 81 331:337

Russell, S; Wefald, E: 1989
Principles of metareasoning

First intl. conf. Principle of knowledge representation and reasoning
Morgan Kaufman
Steel SWD: 1989
Combining probability and epistemic logic
unpublished, dept CS, Essex U
Steel SWD, Reichgelt H: (forthcoming 1990)
Knowing how and finding out
Ninth UK Planning SIG, Nottingham
Tate A: 1977
Generating project networks
IJCAI 77 888:893

Unintelligible formal details. These are included for later reference, rather than because they make free-standing sense.

This is the formal basis on the assumption that
expected utility is a random variable, not a new measure

Quotient structures are not needed. But forming them is possible, and may lead to a Kozen probabilistic DL.

---

states are first-order models	S
sentences are true at states	s |= E
primitive change	Change : S -> S total
times	T
primitive interval	Interval : T -> T total

---

phase space	S x T

analogous to S x T x H, but no apparent use except exposition

---

histories	H
ensemble of MicroState	MicroState =def S x T x H
worldlines	WL : H -> T -> S
microstates where E is true	|| E || =def { p:MicroState | p |= E }

---

language defined at microstate p = <s,t,h>

Expr	|[Expr]| <s,t,h>
atomic sentence E	iff s |= E
E & F (etc)	|[E]| p and |[E]| p

primitive events are partial functions.
(Keep it functional so probability arguments work)
denotation of an event is the same at all points in a frame

Event:MicroState x MicroState

Event ::= Change:S x S (x) Interval:T x T

|[Change (x) Interval]|	{ <<s0,t0,h>,<s1,t1,h>> | |[Change]|(s0) = s1 |[Interval]|(t0) = t1 WL(h)(t0) = s0 WL(h)(t1) = s1 }
E ?	{ <p,p> | p |= E }
Ev1 ; Ev2	|[Ev1]| o |[Ev2]|
Ev1 U Ev2	|[Ev1]| union |[Ev2]|
Ev*	U n (0=<n, |[Ev]| ^ n)
Ev ^ n	|[Ev]| ^ n
	R^0 = Diag R^n+1 = R o (R^n)
|[[Event] E]| p	all p' (<p,p'> e |[Event]| p, p' |= E)

```
                                  |[Event]|(p) defined -> |[Event]|(p) |= E

occ Ev ==def                      - [Ev] #

-----------------------------------------
quotient structures

        x:A / R:A x A  =def  { y | x R y }

        A / R:A x A  =def  { x/R | x:A }

        S:A x A / R:A x A  =def  { <x/R,y/R> | <x,y>:S }

-----------------------------------------
indistinguishability            I: MicroState x MicroState
typically equivalence relation

epistemic state at p: MicroState          p / I

macrostates                     MacroState =def MicroState / I

need not be a partition of ensemble since I need not be an equiv rel

|[ K ]|  = I

K E =def  [K] E

<K> E  ==def  -[K]-E

-------------------------------------
probability

PROB(Set,BaseSet)  =def  m(Set intersect BaseSet) / m(BaseSet)

|[ p(E) ]| p =  PROB( ||E||, p/I )

|[ p(E|F) ]| p =  PROB( ||E||, ||F|| intersect p/I )

m must have Positivity, Sigma-additivity
so that PROB has those + Normalcy, hence Kolmogorov
to get that,

        m: MicroState -> R              m(p) = 1   or some constant

problems with undefinedness

PROB(Set,BaseSet) defined            iff     BaseSet /= {}

prob( E ) is defined            iff     <K> t
prob( E | F ) is defined        iff     <K> F

-----------------------------------------------
expected value

EXP(X,Set) =def Sum p ( p e Set, |[ X ]| p . PROB({p},Set) )

           =    Sum p ( p e Set, |[ X ]| p . m({p}) ) /
                Sum p ( p e Set,             m({p}) )

|[ exp(X) ]| p  =       EXP(X, p/I)

|[ exp(X|E) ]| p  =     EXP(X, ||E|| intersect p/I)

problems with undefinedness

EXP(X,Set) defined              iff     Set /= {}
                                        all p(p e Set, |[ X ]| p defined )

exp( X ) is defined             iff     <K> t
                                        K (      X def )
exp( X | E ) is defined         iff     <K> E
                                        K ( E -> X def )

K T=x  |-  exp(T)=x

-----------------------------------------
utility

goodness of microstate depends only on state

        gs: S -> R

        g: MicroState -> R              g(<s,t,h>) =def  gs(s)


⌊[ $ Event ]| p          =      if      |[ Event ]| p  =  <p,p'>
                                then    g(Event(p)) - g(p)
                                else    undefined
```

```
$A def              iff     occ A

-------------------------------------------
rules to prove
i e I

if      A o I  included in  I o A
then    K [A] E  |-  [A] K E

Ai p.d. = Ai pairwise disjoint =def
x e Ai and x e Aj implies i=j

BASIC $A, assuming that all terms are defined

$E? = 0

$A=x, [A] $B=y  |-  $(A;B)=x+y

if      A, B p.d.
then    $(A U B) = $A  v  $(A U B) = $B

because         occ A^(n+1) -> occ A^n
and             A* = Un (0=<n, A^n)
then            |[ A* ]| is in general a relation
so              $ A* is undefined

possibly an alternative notion of $, eg utility/cycle wd work

OCCURRENCE

E |- occ E?
occ A, [A] occ B |- occ A;B
occ A |- occ A U B              occ B |- occ A U B
|- occ A^0 |- occ A^n |- occ A*

DEFINEDNESS OF $ Event

occ E? |- $ E? def
$A def, [A] $B def |- $(A;B) def
A,B p.d., $A def |- $(A U B) def        A,B p.d., $B def |- $(A U B) def
|- - $(A*) def

ABOUT exp($A), assuming that all terms are defined

K $A=x |- exp($A) = x

|- exp( $E? ) = 0

if      AoI incl in IoA
        exp($A)=exp($A|occ A;B)
        exp($A)=x
        [A] exp($B)=y
then    exp($(A;B))=x+y

if      A, B p.d.
then    exp( $(A U B) ) = exp($A).p(occ A) + exp($B).p(occ B)

if      Ai p.d.
then    exp( $(Ui(Ai)) ) = Sum i( exp($Ai).p(occ Ai) )

|- - exp($A*) def

ABOUT exp($A|.), assuming that all terms are defined

|- exp( $A | E ) = exp( $(E?;A) )

if      Ei & Ej implies i=j
then    exp( $A | Or i(Ei) ) = Sum i( exp($A|Ei) . p(Ei) )


-------------------------------------------
MIXED STRATEGIES
i e I

|[ Ui( Ai:xi ) ]| p   =

        { <p,p'> |      <p,p'> e Ui( Ai ),
                        RAi = Ai restrict p/I,
                        PROB( dom RAi, dom Ui(RAi) ) = xi  }

if RAi are not also pairwise disjoint, then eu arguments will be hard.
but that is separate.

if      occ Ui(Ai)
        p( occ Ai | occ Ui(Ai) ) = xi, for all i
then    occ Ui(Ai:xi)
```

```
if      occ Ui(Ai)
        Ai p.d.
then    exp( $(Ui(Ai:xi)) ) = Sum i( exp($Ai).xi )
```

Current research of Kave Eshgi:

Kave Eshgi
Hewlett Packard Labs., Bristol BS12 6QG

Currently, I am working on the application of model-based diag nosis theory to realistic circuit diagnosis problems. There are two sides to this research:

1. Current model-based inference techniques are hopelessly inefficient when applied to real problems. We are working on the development of new inference techniques and algorithms to make the computations more tractable.
2. The diagnosis framework developed by Reiter assumes that the structure and behaviour of the circuit are described in a monotonic logic language. We have found that in realistic applications, it is useful to have non-monotonic constructs to describe default behaviour. Thus at the theoretical level, we are working on the extension of Reiter's framwork to include non-monotonic constructs in the description language.

We have developed a link between Reiter's diagnosis theory and the stable model semantics of logic programming. This has given us a new perspective on the computational and representational problems mentioned above.

SESSION 6:
DIFFERENT VALUE SYSTEMS

PRESENTED BY: George Kiss

REPORTED BY: Jim Doran

Also reports of current work by:

George Kiss on autonomous agents

Jim Doran on the Tiananmen Square Problem

Different Value Systems

Presenter:George Kiss

Rapporteur: Jim Doran

George began by explaining that in his talk the emphasis would be on the search for the right intuitions and their implementation, as contrasted with formalisation. In his view, formalisation was not a necessary step toward implementation, but rather a parallel activity.

Complex goal directed agents were his focus of interest. They should be intelligent (capable of flexible, adaptive, goal-directed problem-solving behaviour) and autonomous (capable of setting up their own goals based on their own interests and achieving those goals through efficient action).

Desirable characteristics of agents were:

Generality (robustness, flexibility)
Power (economy of resource usage)
Act on and react to environment
Sophisticated interagent interaction (cooperation, competition)
Aware (posessing and using values)

'Folk' psychology provided the following classification of attitudes:

Cognitive: knowledge and belief (epistemic)
Conative : action, wants, intentions
Affective: like, dislike, values

But what was to be done with these intuitive insights? Formalisation had made significant progress only with epistemic -- although there was work on preference systems.

George then turned to physical implementation, specifically the idea of a state based interpretation of knowledge: an epistemic interpretation could be given to a physical machine state, corresponding to a modal logic (cf notably Rosenschein's situated automata theory). Hence a correspondance between a formal domain and a physical device could be established.

But what about using physical concepts to address the dynamics of action? One possibility was to use the concept of state transition diagrams. Actions produce state transitions. The dimensionality of the state space is determined by the action repertoire, and action sequences produce state space trajectories. George also pointed to the potential significance of "attractors and "repellors" in state space (corresponding to achievement and avoidance goals respectively) and how nonlinear systems dynamics theory ('chaos theory') was showing that relatively simple specifications could lead to very complex behaviours involving limit points and similar concepts, for example the work of Hogg at Xerox]

Sam Steel questioned the relationship between the dimensionality of the

the state space and the actions which define transitions in it. George replied that axes are implicitly defined by compositions of basic actions.

George then pointed to the relevance of the Boltzmann machine where energy minima correspond to fixed points which in turn correspond to concepts of interest to a recogniser. But connectionist research had concentrated largely on recognition and learning. Could something similar be done on the action side with the fixed point corresponding to goals? [Later, in the final discussion, George argued that this did not necessarily imply an unchanging set of goals]. There was an interesting distinction to be drawn between local and global limit behaviour which might be related to the relationship between goals and values.

Nigel Seel asked where catastrophe theory fitted in to this. George replied that it had been shown to be a special case of non-linear dynamics.

George then moved on to talk more specifically about implementation ie building an agent. Key issues were:

Power V generality

Vertical layering, for example:

fully deliberative action (Abstract and explicit world representations)

complex skilled routines (Some repns. Some 'concious' decision making)

simple reflex actions (No repns. or 'concious' decision making)

George suggested that symbolic representation is the most general representation system: low levels cannot afford to be symbolic, but must go down to the purely physical. This can be summarised as: "Implicit representation for performance, explicit representation for generality -- the power v generality trade off". BUT (physical) automata are everywhere in the hierarchy in the sense that at all levels there is a hardware that implements everything.

George then described ongoing implementation work in the MMI domain. An agent architecture had been implemented in which a value hierarchy drove the choice of actions in particlular situations. Experiments had been performed in which the explicit value hierarchy (in the form of reduction rules) had been compiled into a run-time m/c at the (emulated) hardware level yielding a speed up of three orders of magnitude. Architectures intermediate between uncompiled and fully compiled were also being explored.

In the **discussion** that followed George's talk **Sam Steel** suggested as a synthesis the idea of a spectrum of architectures defined by increasingly explicit repns. **Jim Doran** suggested that there was a 'dual' to the spectrum (or space) of agent architectures, namely the space of task environments to which particular architectures were appropriate. His note on the 'Tiananmen Square' problem was about this 'dual' relationship.

JED

Value Mechanisms in a Theory of Agents

George Kiss

Human Cognition Research Lab

The Open University

Overview of Themes

Agents as dynamic entities: states and transitions

Epistemic interpretation of states

Actions interpreted as state transitions, behaviour as a trajectory

Topology of the state space; attractors

Conative interpretation of attractors: goals and values

Topological distinction between goals and values: local and global properties of state spaces

Describing agents: in folk psychology, in logic, in terms of mechanisms (implementation)

Hints about the mechanistic description

Some details on implementation technology

Some performance data

What are agents?

Agents are the dynamic entities in the world: they cause events

Agents vary in complexity: thermostats, insects, computers, people

Complex goal-directed agents are the focus of interest:

Intelligent: capable of flexible, adaptive, goal-directed problem-solving behaviour.

Autonomous: set up their own goals based on their interests and achieve these goals through efficient action.

Rational: obey certain intuitively correct constraints on their behaviour.

The position of agent research in AI

Agents are *integrated* AI systems

Agent research draws on AI techniques like search, inference, knowledge representation, vision, language, etc.

Agent research is a superset of AI and forms a new paradigm for doing AI. The nearest field is robotics.

What are the desirable characteristics?

Generality (robustness, flexibility)

Power (efficiency, economy of resource usage like time and space)

Act on and react to the environment

Sophisticated inter-agent interaction (cooperation, competition, communication, etc)

Autonomy (knows what is good and bad, sets up goals and pursues them accordingly)

Reflexiveness (capable of self-representation and understanding and hence self-modification)

The Folk Psychology Description: Attitudes

Three classes of attitudes

Cognitive: to do with knowledge or belief

Conative: to do with action, want, intention

Affective: to do with like, dislike, value

The Formalisation of Folk Psychology in Logic

Formalisation of the attitude of knowledge

- Modal logic: possibility and necessity
- Kripke: Possible worlds semantics
- Hintikka: epistemic logic

Other attitudes? - not much work on formalisation, but philosophical analyses are available:

- Practical reasoning (von Wright, Kenny, etc)
- Preference logics (von Wright, Rescher, Chisholm, etc)
- Intentions (Cohen and Levesque, Bratman)

The Mechanism/Implementation Description - 1

Folk psychology or formalisation are *not enough* either for psychology or for AI: statement of the *physical implementation mechanisms* is needed!

The state-based interpretation of knowledge:

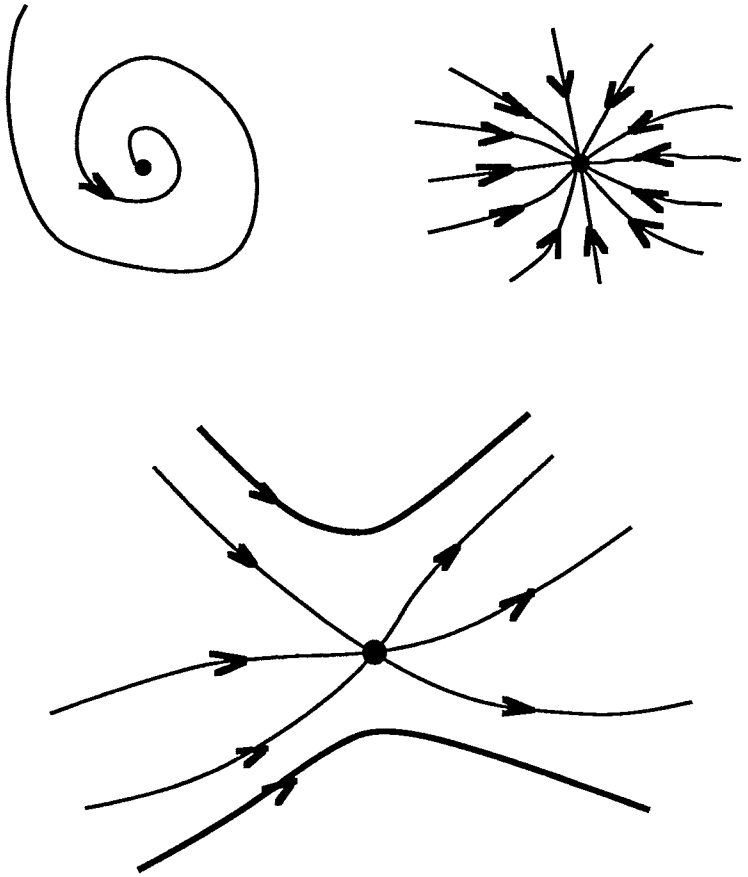
An epistemic interpretation can be given to physical machine states, following Rosenschein. If an internal machine state is reliably correlated with a world state, then the machine can be said to *know* that world state. The world state is decribed by a proposition. The denotation of the machine state is the proposition.

The Dynamics: State Transitions

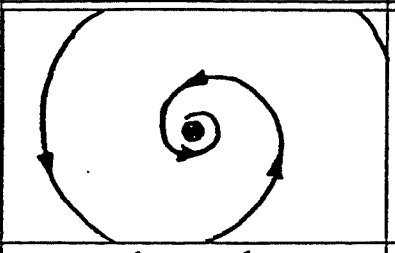
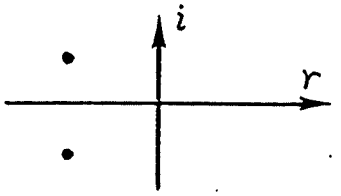
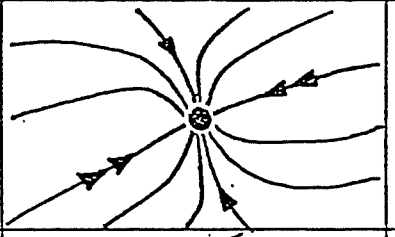
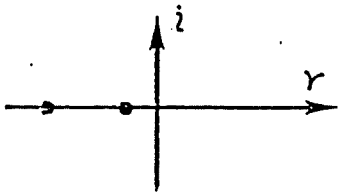
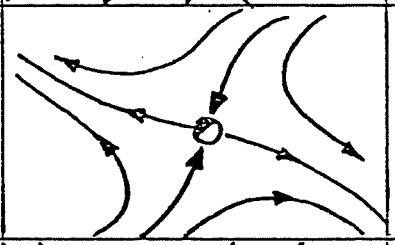
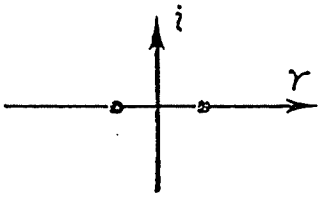
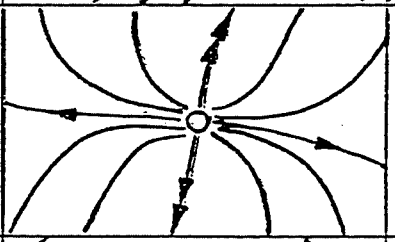
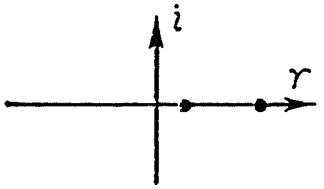
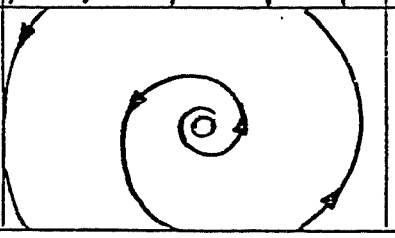
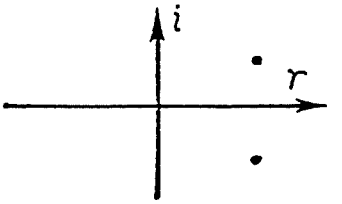
Actions produce state transitions. The dimensionality of the state space is determined by the action repertoire. Action sequences produce *trajectories* in the state space.

Attractors and Repellers in the State Space. The state space has a topology determined by the state-transition mappings corresponding to the actions. Trajectories move towards attractors and away from repellers.

Classification of Attractors in terms of trajectories. Fixed points; saddle points; cyclic attractors, etc. Local and global attractors.

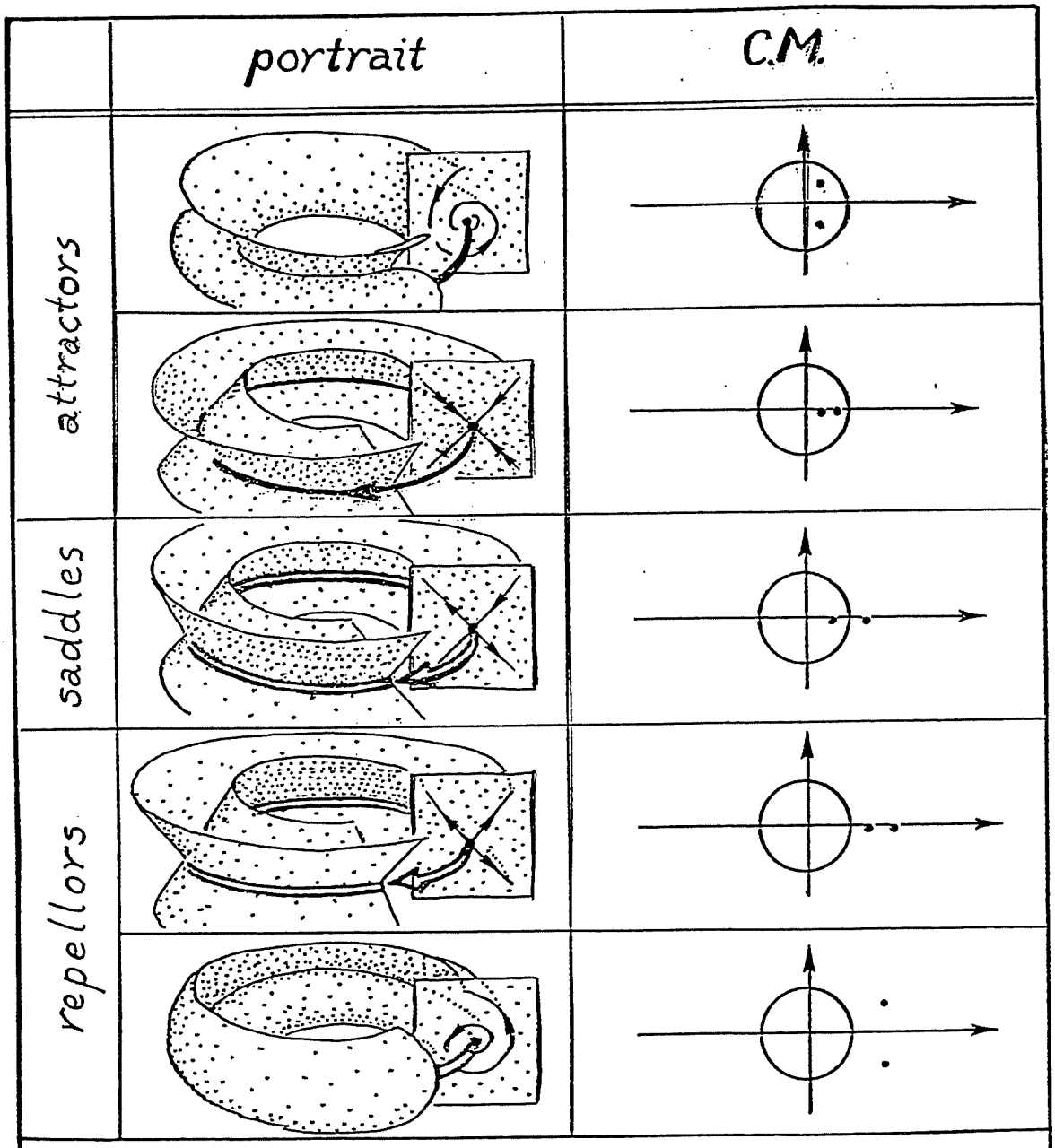


See e.g: DEVANEY, An Intro to Chaotic Dynamical Systems
Addison Wesley 1989

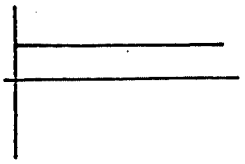
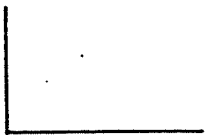
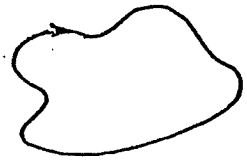
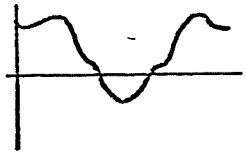
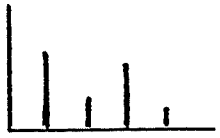
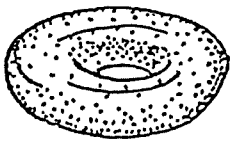
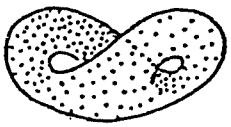
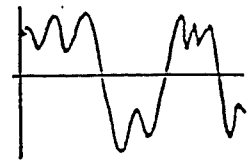
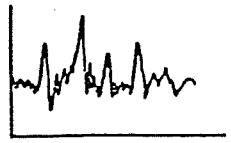
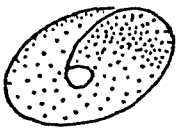
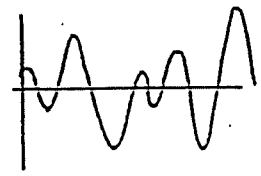
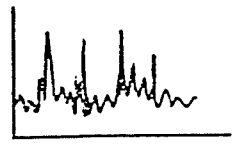
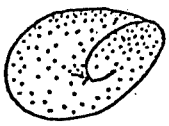
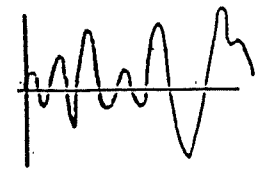
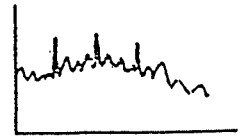
type	index	portrait	C.E.
attractors	0		
	0		
saddle	1		
repellors	2		
	2		

1.4.8 The typical hyperbolic limit points, their CEs, and indices, are summarized in this table.

Actually, two cases are omitted from the table. These are the hyperbolic attractor and repellor with equal (real) CEs. They are classed among the degenerate cases, even though they are hyperbolic, because they are transitional phenomena between the nodal and spiral types. The cases shown are all the *elementary* ones, meaning hyperbolic, with distinct CEs.



2.5.7. A limit cycle is called *elementary* if it is hyperbolic, and its CMs are distinct (no two equal). All the elementary limit cycles in three space are summarized in this table.

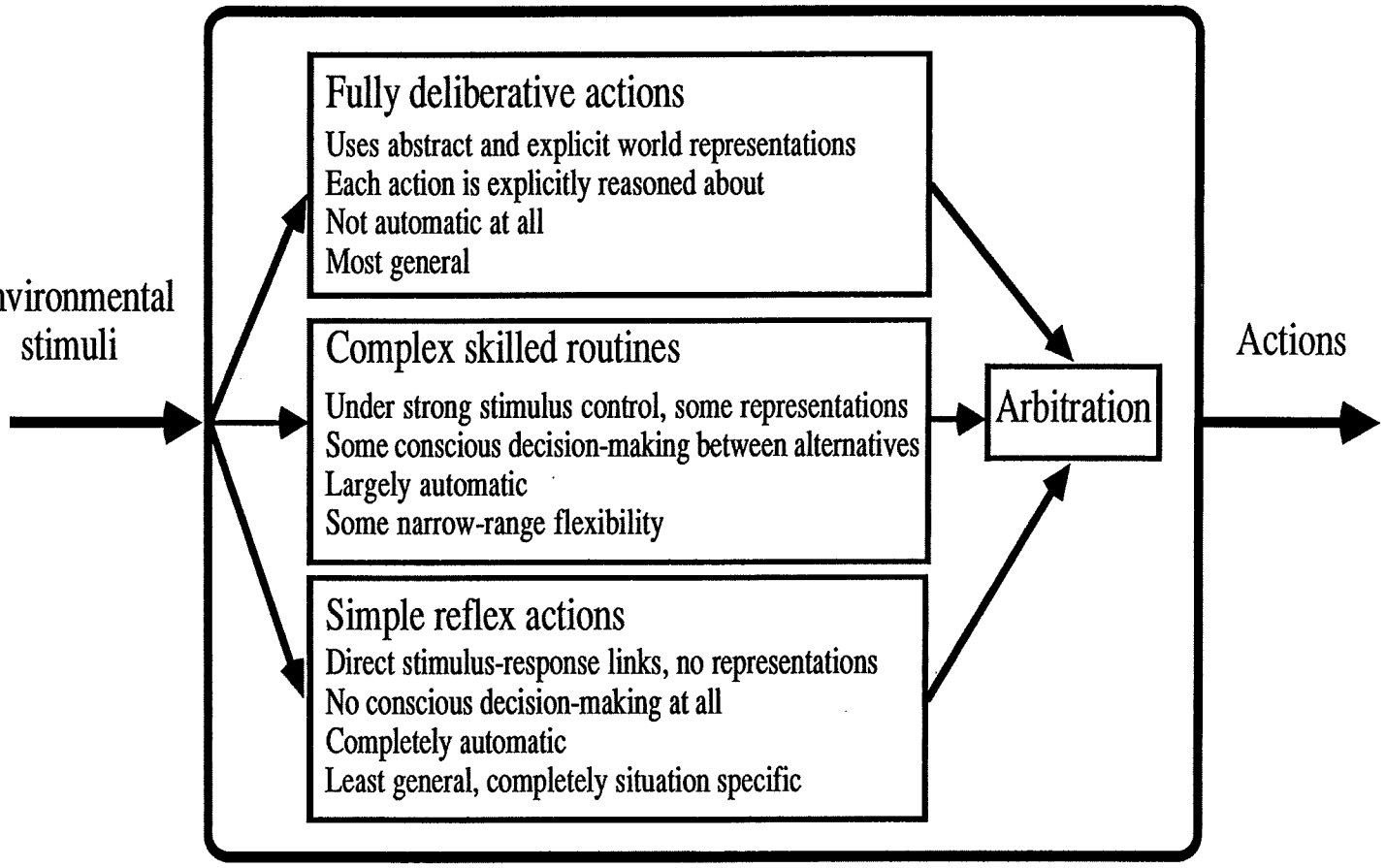
NAME	PORTRAIT	TIME SERIES	SPECTRUM
point			
closed orbit			
Birkhof Bagel			
Lorenz Mask			
Rössler Band			
Rössler Funnel			

4.5.7. Here is a summary table of the exemplary attractors we have presented, with sketches of their characteristic output. One could wish for an extension of this table, showing all possible attractors likely to arise in experiments and applications. But at this point, that's a big wish.

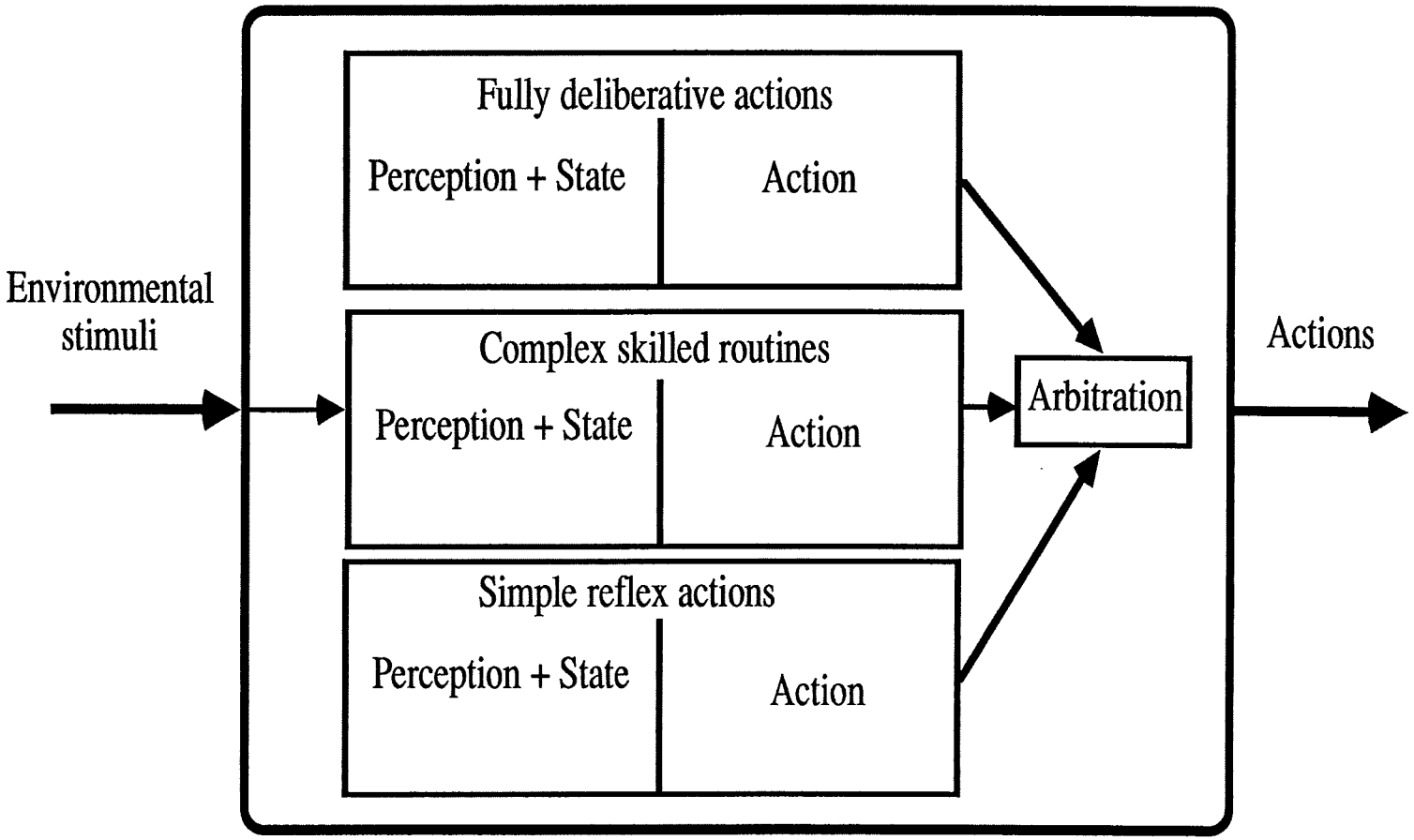
The Mechanism/Implementation Description - 2

Architecture Issues. Modularisation

Power vs Generality: Vertical Layering



Perception vs Action: Horizontal Layering



The Top Level of the Architecture

All actions at this level of the architecture are basic agent actions. These are the simplest actions the *agent* does (as opposed to a process in some "low-level part" of the agent)

Full deliberative reasoning about what actions would produce maximal utility in the current situation.

Continually attempts to decide:
"What is the best action I can do here and now, given my long-term aims?"

Having derived a partially ordered set of actions, commit the agent to one of them.

Value Mechanisms in Agent Theory

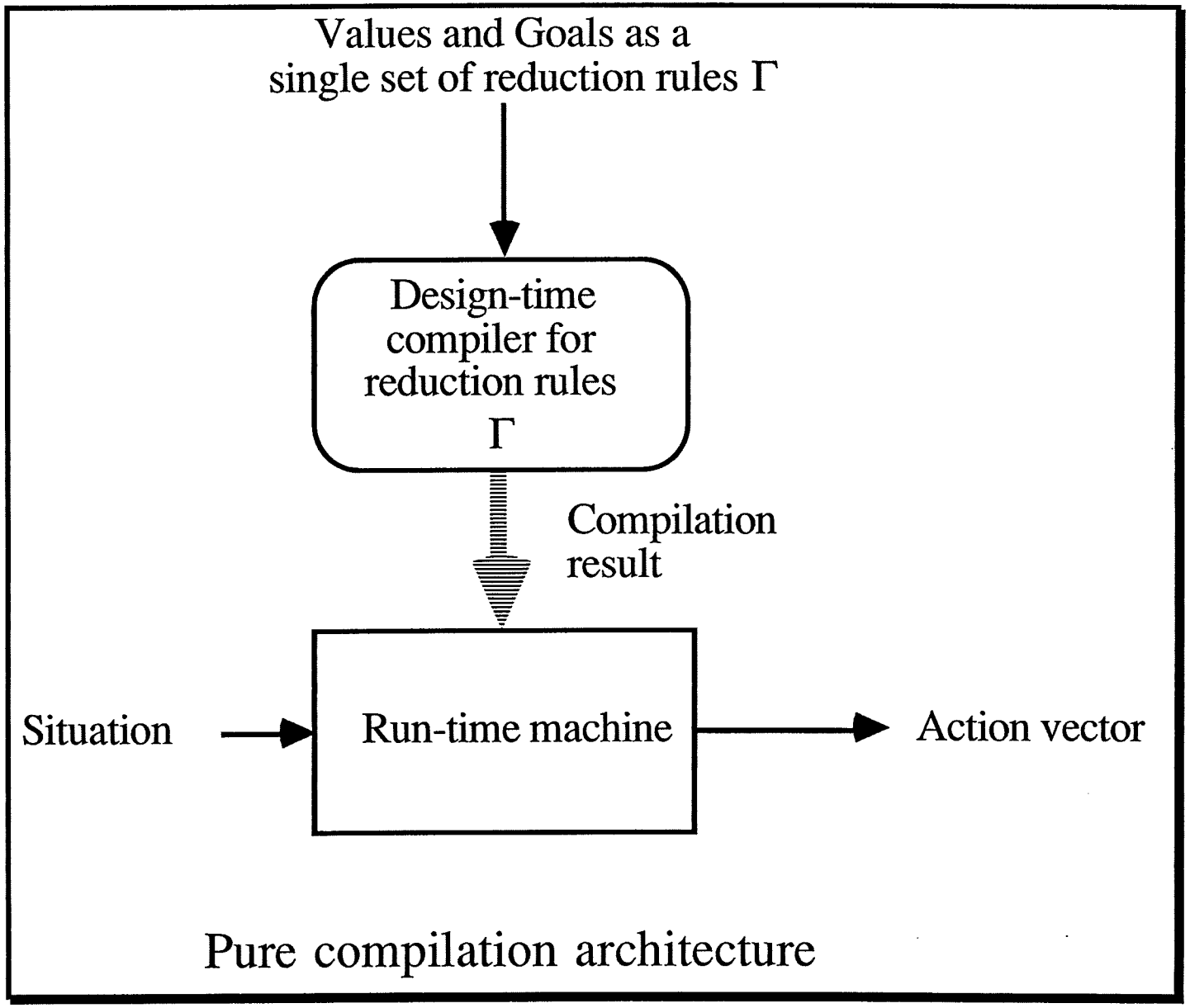
The behaviour of the agent is determined by its value system from which it derives goals.

The *value system* describes the *long-term global attractors* of the agent. Moving the situation nearer to an attractor produces satisfaction.

Goals are situations that can be achieved in the *short-term, locally* (in a low-dimensional sub-space of the global state space).

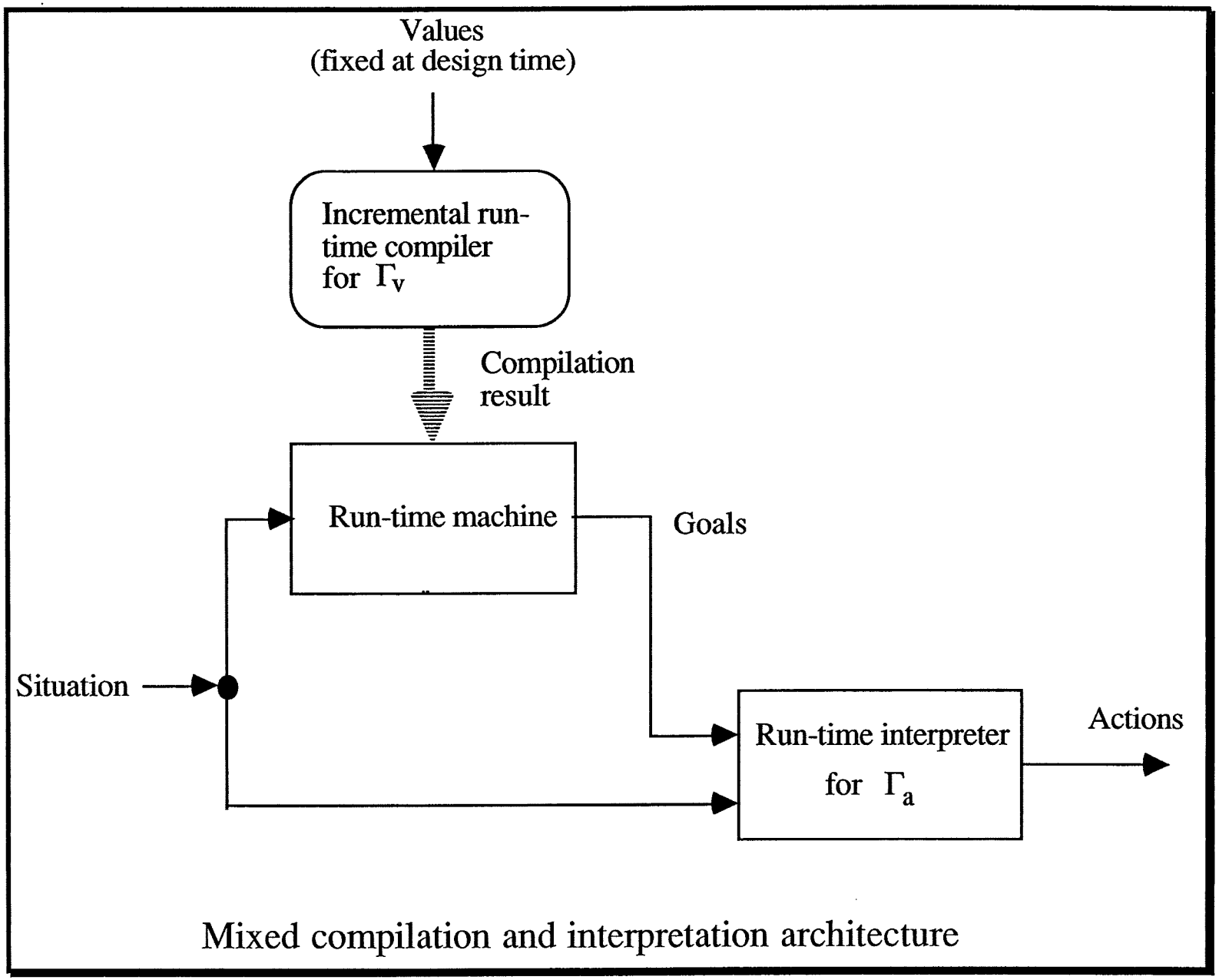
Values and goals are represented *explicitly* at higher vertical layers of the architecture, but only *implicitly* at the lower layers.

Implicit representation means that goals and values are compiled into a run-time machine, using a set of goal-reduction rules Γ. The run-time machine maps situations into actions and shows goal directed behaviour, but it has no explicit representation of goals.


Values and Goals as a
single set of reduction rules Γ
Design-time
compiler for
reduction rules
Γ
Compilation
result
Situation
Run-time machine
Action vector
Pure compilation architecture

Explicit representation of goals means that the situation-action mapping is interpretively produced at run time, rather than compiled at design time.

The values may be still represented only implicitly by using a design-time compiler, or a run-time incremental compiler, to permit run-time acquisition of reduction rules in Γ_V.



Mixed compilation and interpretation architecture

An Illustrative HCI Application Domain Scenario:

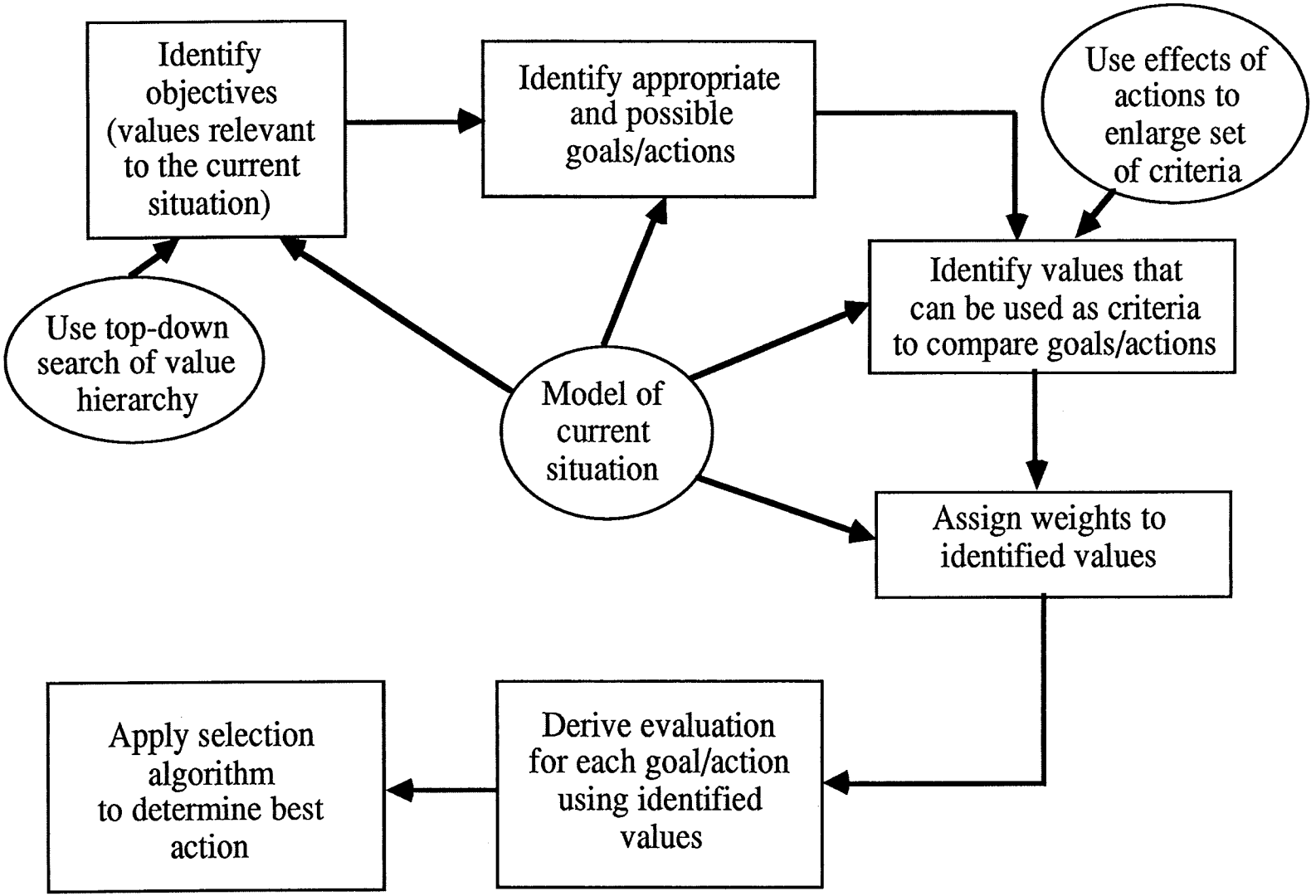
A computer user tries to delete the kernel file of the operating system.

The system's response depends on the situation.

If the user is a novice, it may refuse to delete the file and/or initiate a dialogue.

If he is experienced, it may delete the file.

The Action Selection Mechanism at the Top Level



Notes on the Action Selection Mechanism

Possible mechanisms to detect the relevance of situations to values are: activation; pattern matching; ... ; full reasoning. Reasoning is the most general mechanism.

Relevant values are activated at the lowest (most specific) level of the value tree, i.e. the leaf nodes.

Actions or goals which would lie on the state-space trajectory towards the satisfaction of the value are identified, again by a range of possible relevance mechanisms. The most general one is again reasoning about means-ends relationships.

Weights are assigned by modulating long-term permanent weights

Utilities of candidate actions are computed by summing over active leaf nodes and multiplying the sum by the curent leaf node weight.

The result is an ordered list of candidate actions. The selection algorithm can take the highest ranking action, or possibly apply other criteria to override the ordering.

Experimental Implementation

A common LISP program that implements the action selection process.

The value tree is the central data structure.

Reasoning is implemented through a forward chaining rule interpreter, but other inference mechanisms could just as well be used.

The program has been tried on the initial situation of the scenario.

Running time for one decision is of the *order of minutes* on a Mac II.

Program size is 300k, not including LISP run-time system of about 800k.

Action Selection at Low levels

The goal-reduction rule set Γ:

```
(defgoalr (maint survival)
  (prio-and (maint system-running)
            (maint user-satisfied)))

(defgoalr (maint system-running)
  (maint system-files-available))

(defgoalr (maint system-files-available)
  (maint (present-on-disk *kf*)))

(defgoalr (ach (make-present-on-disk *kf*))
  (if (know-kf-on-disk)
    (do anything)
    (do restore !*kf*)))

(defgoalr (maint (present-on-disk *kf*))
  (ach (make-present-on-disk *kf*)))

(defgoalr (maint user-satisfied)
  (prio-and (maint no-info-loss)
            (ach assist-user-goals)))

(defgoalr (ach assist-user-goals)
  (or (ach execute-user-command)
      (ach execute-user-plan)))

(defgoalr (maint no-info-loss)
  (if (andm (know-user-command-received)
            (know-destructive-user-command))
    (and (do refuse (user-command))
         (do ask-user !*why-do-you-want-to-do-
         this?*)
         (do listen-for-answer !*listen*))
    (do anything)))

(defgoalr (ach execute-user-command)
  (if (andm (know-user-command-received)
            (know-destructive-user-command))
    (not (do execute (user-command)))
    (if (know-user-command-received)
      (do execute (user-command)))))

(defgoalr (ach execute-user-plan)
  (if (equalm (user-plan) !0)
    (and (do ask-user !*what-is-your-plan?*)
         (do listen-for-answer !*listen*))
    (do execute (user-plan))))
```

Experimental Implementation

The reduction rules are compiled into a low-level machine of approximately 100 primitive elements.

This machine is then simulated either in MC680x0 machine code, or in C or in Common LISP, or else it could be wired up in hardware.

Running time for the same decision is of the order of a few hundred milliseconds in the Common LISP simulation on a Mac II or in the C-compiled simulation on a Sun3.

Thus speedups of three or four orders of magnitude could be achieved by hardware implementations, compared to rule-interpreter technology.

Program size is about 10k bytes from C compilation, not including run-time libraries, amounting to 100k.

Current research of George Kiss: Research on Autonomous Agents

George Kiss
HCRL, Open University

High-Level Dialogue in Man-Machine Interaction. This is an Alvey project in collaboration with British Telecom, with a total funding of **£250,000** over three years. The research investigates HCI dialogue as an interaction between autonomous agents. Dialogue is regarded as a special case of general agent action aimed at other agents. The project is mainly concerned with the development of an appropriate theory, while also engaging in the construction of illustrative software to demonstrate certain concepts from the theory. The project has so far concentrated on the use of axiological (value-related) attitudes in designing and implementing interactive systems.

Theory of Autonomous Agents. This research is concerned with the development of a theory of natural and artificial agents that are capable of autonomous actions through which they pursue their interests. Focal topics of the research are the epistemic (knowledge, belief), praxiological (want, intention, volition) and axiological (like, dislike, value, preference) attitudes agents may have towards the world; the concepts of self and commitment to an attitude; the distinction between basic, reflex and fully deliberative actions; the relationship between agent-theoretic concepts and mathematical system dynamics.

Implementation Architectures for Autonomous Agents. This research investigates implementation mechanisms for various concepts used in characterising agents and the way in which such mechanisms can be integrated into a unified architecture. Focal topics of the work are: power versus generality tradeoff; hierarchical modularised organisation; the role of distributed system concepts; state-based implementation of epistemic attitudes; using functional, object-oriented and agent-oriented programming languages for implementation work.

Current research of Jim Doran: The Tiananmen Square Problem

Jim Doran
Dept. of Computer Science, University of Essex

Original version October 1989
This version January 1990

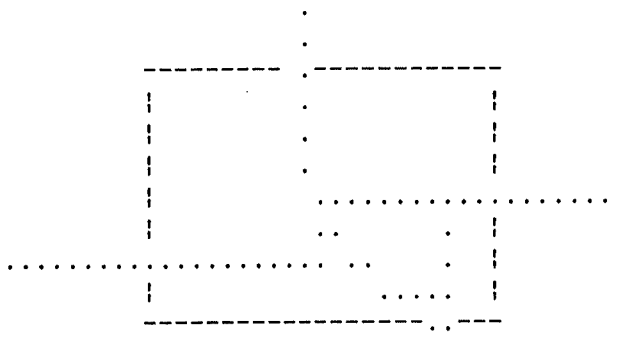
The META PROBLEM

Pose a precise non-trivial multiple agent coordination problem to which a dynamic organisation with transient participants is the only solution

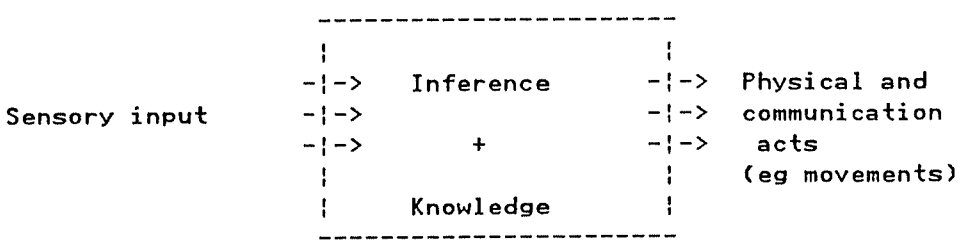
We approach this meta-problem by noting that in everyday experience traffic systems (air, road, sea etc) require organisation. Absense of organisation is only acceptable at very low traffic densities. Hence:

The TIANANMEN SQUARE PROBLEM (TiSP)

The following diagram represents a large parade ground or 'square' across which vehicles move freely to achieve transits between exits



The following diagram represents an autonomous vehicle (AV)



The TiSP is to determine what should go 'inside' the AV (at the level of AI processes) so that it will be able to perform certain tasks (square crossings) as to be specified.

Distinguish:

Design specification time
(When we specify the design of the AV(s))

Task specification time
(When tasks are specified to an AV)

Action time
(When an AV executes an action)

Version 0

A rectangle (the square!) has a number of entry/exit points (ee-pts).
An AV exists.

At design time, we, the designers of the AV, know the dimensions of the rectangle and the locations of the ee-pts .

The AV must be able to achieve any task
(given at task specification time) of the form:

You are at ?ee-pt1, travel from ?ee-pt1 to ?ee-pt2
by as short a route as you can.
(A route is travelled by a sequence of directly executable actions).

The AV is able to execute directly (and exactly) actions of the form:

Move for distance X on bearing Y

THE PROBLEM is to define the AV's processing.

COMMENT

The problem is trivial. We provide the AV with a
list of all ee-pts and their coordinates, and an algorithm
which, given values of ?ee-pt1 and ?ee-pt2 computes the required
distance and bearing and executes the corresponding action.

Version 1

Now suppose that immobile rectangular obstacles exist within the
given rectangle, whose size and locations are known to
us at design time. All else as in Version 0.

There are now (at least) three different designs which
we may build into the AV at design specification time:

(1) Compute routes at design time for all possible ee-pt pairs,
and provide the AV with a table of them. Set it to
look up a route as required at task specification time
and to then execute it.

(2) Provide the AV with a complete map of the 'square'

and the A* algorithm and set it to compute and then execute routes as required at task specification time.

(3) Make it reactive, that is, ASSUMING that short range sensing is available for the AV, arrange for it, for example, to proceed directly towards its goal if there is no obstacle immediately in the way, but to track round the perimeter of the blocking obstacle as long as this requirement remains unsatisfied.

Note that (3) requires short range sensing. Until now, sensing has not been assumed/required.

Variations on these 3 possibilities are possible.

Version 2

Now suppose that **mobile** rectangular obstacles exist within the given rectangle. All that we know of these at design time is that they exist. All else as in Version 0.

Much now depends upon how much information is available to the AV (ie how much it can sense) at task specification time. Two broad possibilities are:

(1) Design the AV so that after task execution time it alternates between a route planning stage (involving A*) and route execution, using additional information derived from sensors as it becomes available.

(2) Design the AV to react to local information holding to a bearing calculated at task specification time. It seems unlikely that success can be guuarranteed for every particular task.

Version 3 -- This version is the 'full' TiSP

We now suppose that the requirement at design time is to design *a very large number* of AVs (not necessarily all with the same design) so that as they execute tasks a cost function is minimised. The cost function expresses (a) a (prohibitive) penalty for any type of vehicle collision, and (b) a penalty for any vehicle excess travel time (ie longer than the minimum 'as the crow flies' time).

The new aspect is the requirement to avoid collisions (either between vehicles or between a vehicle and an obstacle) and the costs associated with them.

Much depends upon the AVs' ability to detect other vehicles in their locality and to change speed (eg stop) quickly. But note that even with good sensing and fast reactions, there is still the risk of 'turning into' a collision. A further major consideration is the possibility of 'inaccurate' execution.

Version 3.1

Assume:

-- planar formulation

-- that the 'square' has dimensions 10000 x 5000

-- that 10 ee-pts each with width 5 are distributed at random round the square (uniform probability distribution excluding corners and overlapping).

-- that obstacles may be rectangles of any size up to 500 x 500 and move as follows -- at each time unit either no motion or one unit N,E,S or W with probability 0.01 .

-- that each AV is a disk of radius 0.5 and can rotate on its axis

-- that in each time unit a random number of AVs (between 0 and 5) present at each ee-pt with destination points draw at random from those available. (Uniform probability distribution in both cases).

-- that the speed range of AVs (both backwards and forwards) is 0-100 per time unit.

-- that AVs execute actions perfectly (in the absense of collisions).

-- that AVs have sensing (perfect) up to a distance of 10

-- that the computations performed by AVs take no time

-- that an AV can store an infinite amount of information.

and that all the above is known to us at design time.

Assume also that AVs can communicate without error and without information limit over a range of up to 100, but with a time lapse (irrespective of message size and distance) of 5.

Comment: it would, of course, be possible and conventional to formulate the problem with variable parameters rather than specific figures (eg for the maximum speed of the AVs). To do so would, however, add much complexity as well as generality to the problem since the solution (below) would certainly vary with alternative parameter combinations.

What designs should be provided for the vehicles if collisions are to be avoided and if the mean excess journey time across the square is to be minimised?

Does the answer to this question involve self organisation?

Can it be answered other than by computer-based trial-and-error experimentation?

SESSION 7:
CHANNELS FOR DIALOGUE

PRESENTED BY: Phil Stenton

REPORTED BY: Ann Blandford

Also reports of current work by:

Phil Stenton on designing cooperative interfaces

Ann Blandford on a model of tutorial dialogue

Channels for Dialogue

Presenter:Phil Stenton

Rapporteur: Ann Blandford

Phil gave a resumé of work going on at Hewlett Packard involving the development of agents. The work focuses around the development of interfaces which allow the user to more accurately express what they are trying to achieve when using an application package, such as a database, and respond more appropriately to the user's expectation.

'Wizard of Oz' studies have been done, in which the user believed that they were querying a natural language (NL) database when in fact they were communicating with a human 'wizard' who had hardcopy of the data, and typed appropriate responses back. The protocols taken could be analysed in terms of both strategy (what the user was trying to find out) and process (how they went about it). NL technology is not yet capable of dealing with extended dialogues of the type generated in this study, so a windows system has been developed based on the results, as an interface to an expenses monitoring database. The 'buttons' available in the interface (i.e. the clickable boxes) indicate the extent of the capability of the system to the user (often a problem with NL interfaces, where the user does not know what the system can and cannot deal with).

The intention now is to 'ramp up' the capability of the interface, introducing 'agents' as interfaces to the system, or network, rather than just as front-ends on individual application programs. This involves the development of distributed agents, some acting as 'personal assistants' to users, others dedicated to particular tasks, which can communicate with each other appropriately.

Phil showed a 15-minute video, outlining HP's vision of the state of the art in 1995. As well as advances in conventional technology (such as comms), this showed a vision of agents, characterised as cute little robots on the screen, which could:

- remind the user of appointments
- generate synthesised speech and understand spoken commands (!!!)
- write reports, giving recommendations, justifications, sources and logic
- get information from other sources (not just HP systems)
- process data in defined ways (e.g. "run the numbers through 'Finance'")
- filter news from national news sources (including TV??) which was likely to be relevant to the company
- communicate with other agents
- deal with aspects of computer security

Following the video, Nigel Seel queried the notion of 'agent' as portrayed in it; was the agent anything more than a more advanced i/o mechanism? It was agreed that the only instance of the video agent showing any obvious intelligence was in its presentation of its rationale when making recommendations in a report. Otherwise, the video promoted the view that all that was important was information (and having sharp blue eyes and chiselled features!!). HP customers apparently like the video more than recalcitrant academics do! It was felt that to call something an 'agent' it had to be capable of more than simple resource management; for example, it might be able to participate in cooperative problem solving. Nigel Shadbolt gave the example of

a knowledge source which knew how to apply repertory grids (i.e. had expertise) and suggested that this, through having more 'depth' might qualify as an 'agent'.

George Kiss raised the question of how agents might best be employed. He observed that they are needed for the little problems as well as the big ones (though everyone else agreed that the example he gave to illustrate this point - involving finding the source of error when a porting exercise failed - counted as a big problem, not a little one!) This example led to the suggestion that there would have to be 'clerk' agents which did routine work, and 'mechanic' agents which found and fixed errors.

David Connah observed that in order to be acceptable to users, agents would have to be more than simply competent; they should be able to deal with situations much more complex than those they generally encountered. Others disagreed, observing that most people already display too much trust in non-intelligent computer technology. Steve Pullman considered the interesting bit of the video to be the agent's ability to justify its behaviour, and suggested that this should be the basis for sociological trust - i.e. the agent must be accountable for its actions. This necessitates the sort of 'heavy duty' agents David proposes.

Nigel Shadbolt noted that two important questions raised by the video were that of where delegation (from one agent to another) and search (for information) stop in a DAI system, and how an agent would prioritise its commitments.

These issues were not discussed further, as time did not permit. Maybe next year...

Putting Agents to work whilst avoiding the 'Donkey'

Phil Stenton
Advanced Information Management Dept.
Hewlett-Packard Laboratories

Abstract

This presentation discusses our work on the design of cooperative interfaces to information management systems. We have analysed real data from experiments and field study transcripts. We have used the resulting dialogue theories to match interface technologies to dialogue requirements. To test our theories we have designed and built a working prototype in a real domain. We aim to adopt the same approach to devoloping agent architectures. Our interest in agents is to use them as an architectural home for our work on dialogue. Here we present our vision and finish by posing questions to those working on agent architectures.

1 Introduction

The first half of this presentation will briefly describe our work on dialogue modelling. This will provide an historical perspective and explain our motivation for working on agent architectures. The second half of the talk will describe our current work and our vision for the future.

The primary focus for our work has been to develop dialogue systems which can support cooperative problem solving between people and computer software. Current work on cooperative problem solving has three main foci:

1.1 Cognitive Load Distribution

The first focus stems from the observation that 'Expert Systems' do all the problem solving and the user provides the data, whereas information systems provide all the data and the user does all the problem solving (Kidd 1984,85 Allport 1989 Miller 1984, Cohen & Levesque 1987, Allen & Perrault 1980). The work from this camp is aimed at redistributing the problem solving to somewhere in the middle.

1.2 Dialogue Control

The second focus is similar to the first but the emphasis is on control of the interaction. The two ends of the cooperativity spectrum are system controlled and user controlled. The key to cooperativity is the distribution of dialogue initiative (Frohlich et al 1987, Gilbert et al 1987, Whittaker & Stenton 1988 Matthews 1985). Again the goal is to build systems which are positioned somewhere in the middle, customised to the needs of the problem solving task.

1.3 Channel Tailoring

The final focus centres on the design of the communication channel to make the information exchange as easy as possible. Cooperative behaviour is thus supported through the surface characteristics of the interface (Sneiderman 1986, Whittaker & Stenton 1989 Gross 1977 Gross & Sidner 1987, Walker 1989). In this group we include the NL discourse work.

2 Our work so far

In the process of our research we have worked on each of the three foci and collected dialogue

data along the way. Our main (and current) contributions have been in the last group, Channel Tailoring. Our goal has been to design interfaces to information systems which support the kinds of extended dialogue users WOULD LIKE TO HAVE. To this end we have collected experimental data using the Wizard of Oz technique. We have analysed live transcripts of users of an existing information systems solving real problems. Our target user group is business executives and our task focus is marketing and sales information management. The resulting dialogue theories have been demonstrated in a lab demonstrator and a working protoype.

3 Introducing the Agent metaphor

Tailoring the communication channel to support cooperative communication is stretching the currently popular deskstop metaphor beyond its limits. The propogation of intuition afforded by this metaphor is already meeting inconsistencies. These usually appear where the features of electronic storage and manipulation are exploited (e.g copying documents by pointing at them while holding down a combination of keys).

We have been working on the notion that the desktop metaphor is limited but irreplaceable. The solution then is to mix-metaphors by design. That is to find a suitable metophor to facilitate the intuitions of the users as the desktop does for simple paper shuffling tasks. The Agent is the best metaphor around. The goal of our work is to understand what we would want the agent to do before we design the architecture. This is in contrast with an approach which characterises the notion of agency and proposes an architecture to cover its many facets. The role of an agent in our model is that of an 'assistant' (Kiss 1987) where the agent is the interface to a distributed environment providing the tools necessary to for the user to perform tasks. Agent tasks as currently implemented in HP are ojects which store procedural instructions and decisions written in a Task Language which the agent is to execute. Users can write these instructions directly or through a natural language interface or through a learn-by-example procedure.

The development of agent capabilities is to be an incremental one through increasing levels of task specification from procedural to problem specification. It is envisaged that agents will eventually perform tasks across software applications by generating plans from higher level goals given by the user and a knowledge of the software tools available.

4 Questions

During the workshop 'Principled' and 'Pragmatic' approaches was mentioned along with the notion of 'Superbelievers'. Questions for those working on agent architectures are therefore:

- What does it mean to have a principled agent architecture ?
- Where do we get the principles from ?
- Are we modelling the right things and doing justice to them ?
- What is the motivation behind much of the work on Agency: Skill modelling or Psychology ?

5 Conclusions

Work on cooperative problem solving systems can be described as having three broad foci. We are currently working on an agent metaphor as the repository for our dialogue theories. We aim to develop agents as an incremental process of increasing capabilities. Our vision of the future agent is in the role of assistent to help the user navigate and operate in a distributed software environment.

6 References

Allport D. (1989) A Computational Architecture for cooperative systems, Invited Paper to appear in *Proceedings of KBCS Conference*, Bombay, India

Frohlich D.M., Crossfield L.P. & Gilbert G.N. (1985) Requirements for an intelligent form-filling interface. In P.Johnson and S. Cook (Eds.) *People and computers: designing the interface.* Cambridge Univ. Press

Gilbert G.N., Luff P., Crossfield L.P. & Frohlich D.M. (1987) A mixed initiative interface for expert systems: The Forms helper. *International Journal of Man-Machine Studies*, forthcoming.

Grosz B.J.& Candace L. Sidner. (1986) Attentions, intentions and the structure of discourse. *Computational Linguistics* 12 pp. 175–204

Kiss, G (1987)Why bother about goals ? *2nd Intelligent Interfaces Meeting (IISIG)* London pp 66-77 (Invited talk)

Shneiderman B. *Software Psychology.* Winthrop Publishers Inc.

Walker M.A. (1990) Natural Language in a desktop environment To appear in the *Proceedings of HCI International* Boston Sept,1989

Whittaker S.J. & Stenton S.P. (1988) Cues and control in expert-client dialogues *Proceedings of the 26th ACL* pp 123-130

Whittaker S.J. & Stenton S.P. (1989) User studies and the design of Natural Language systems *Proceedings of the 4th EACL* pp 116-122

PUTTING AGENTS TO WORK WHILST AVOIDING THE 'DONKEY'

PHIL STENTON : HEWLETT-PACKARD LABORATORIES

* HISTORICAL PERSPECTIVE
* MOTIVATION
* AGENTS AT HP
* VISION FY95
* QUESTIONS

CHARACTERISING DIALOGUES

* 'CUES AND CONTROL' - ACL '88
* WIZARD OF OZ - EACL '89
* TRANSCRIPT ANALYSIS
* TRANSACTION LOGS AND INTERVIEWS

"KNOW what the Agent has to do before designing the architecture" — Nigel Seel

AGENTS IN A DISTRIBUTED ENVIRONMENT

WORLD:

- INCOMPLETE INFORMATION
- DYNAMIC
- INCLUDES OTHER AGENTS

AGENT:

- PERSONAL ASSISTANT
- MUTUAL KNOWLEDGE OF CONVERSATIONS
- PARTIAL KNOWLEDGE OF S/W WORLD OF NETWORKED ENVIRONMENT

AGENT VISION

Personal Agent:

An intelligent facility that works on behalf of the user and cooperates with other agents to perform tasks within and across objects.

System Agent:

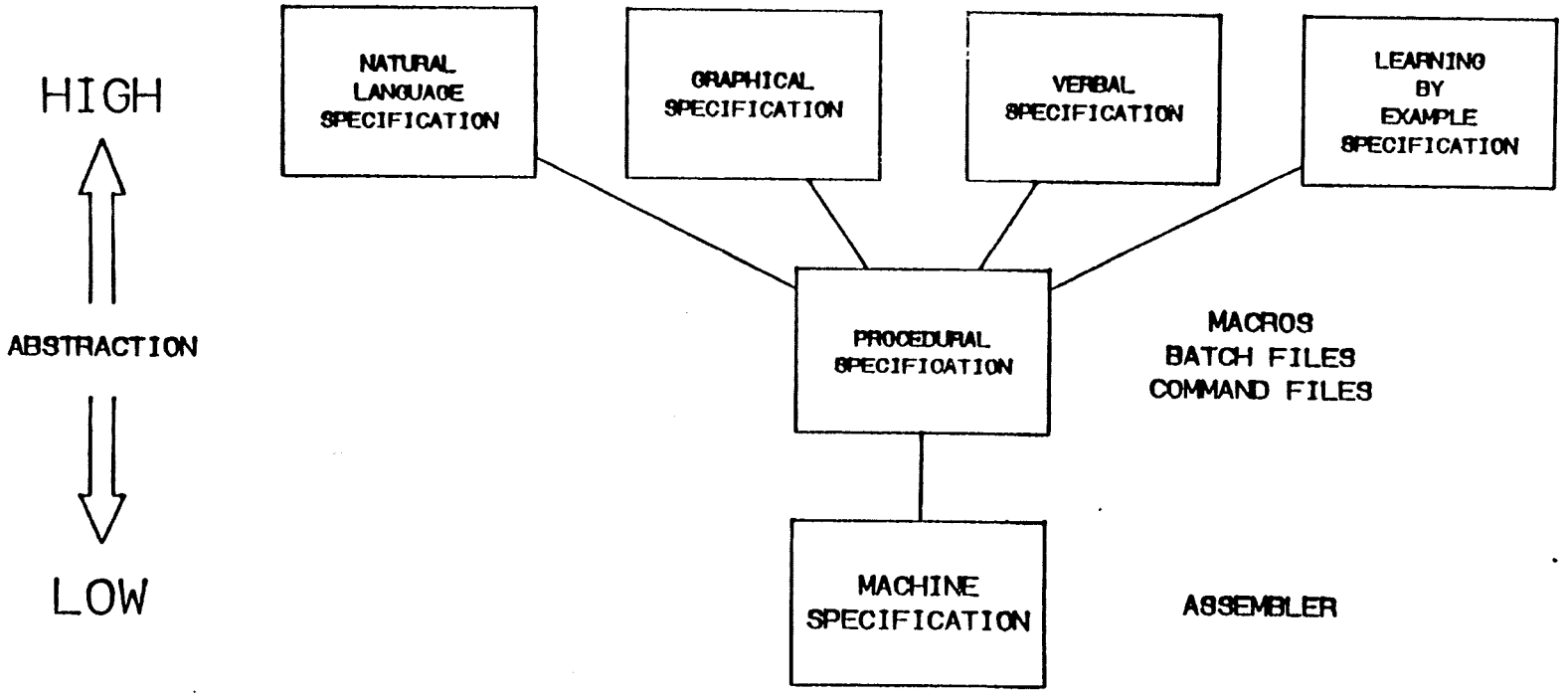
An intelligent facility resident in a network of computer systems that works on behalf of the users and system managers to monitor and control objects.

Think of an Agent as an Assistant or Software Robot.

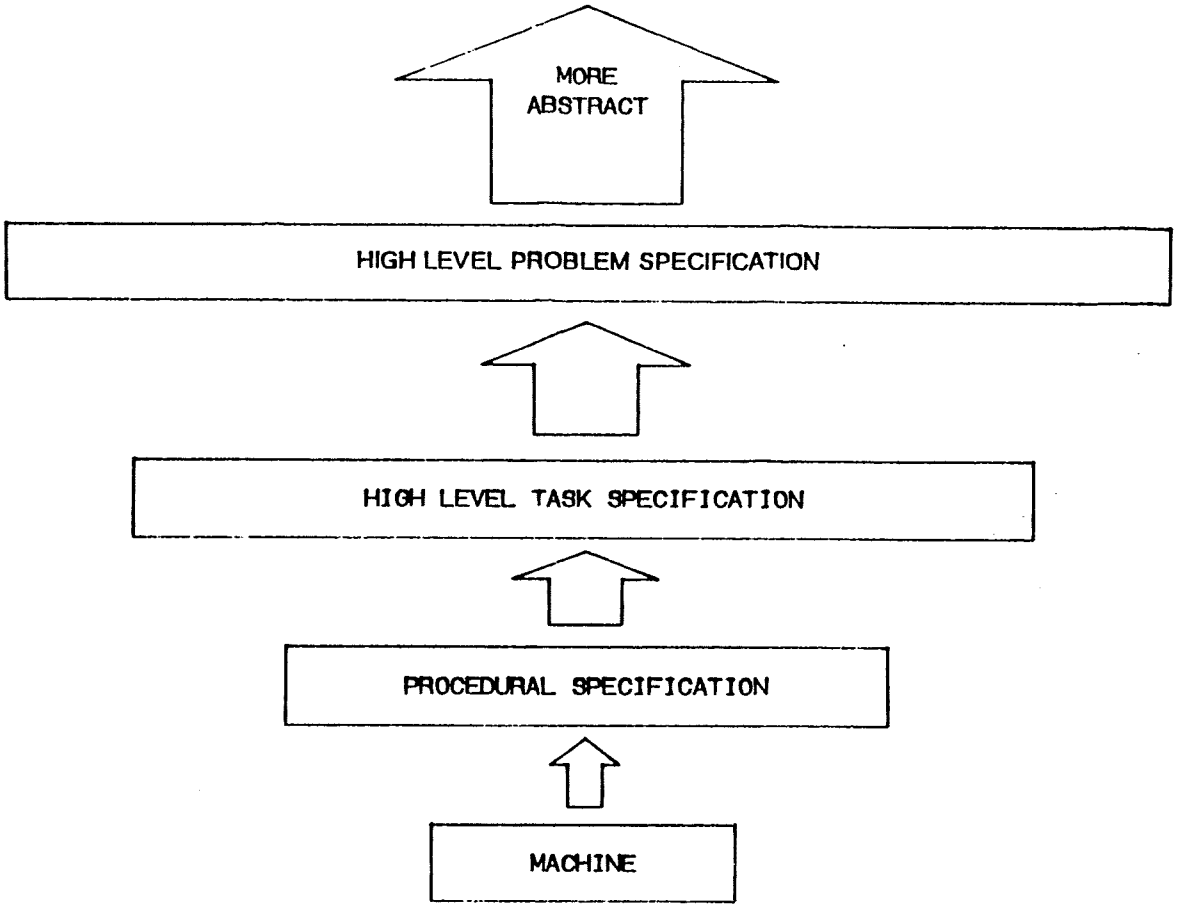
WHAT IS AN AGENT TASK?

An object which stores, and allows the user to edit, procedural instructions and decisions written in Task Language, which the Agent is to execute.

HIGH LEVEL TASK SPECIFICATION



AGENTS SHOULD SUPPORT ABSTRACTION



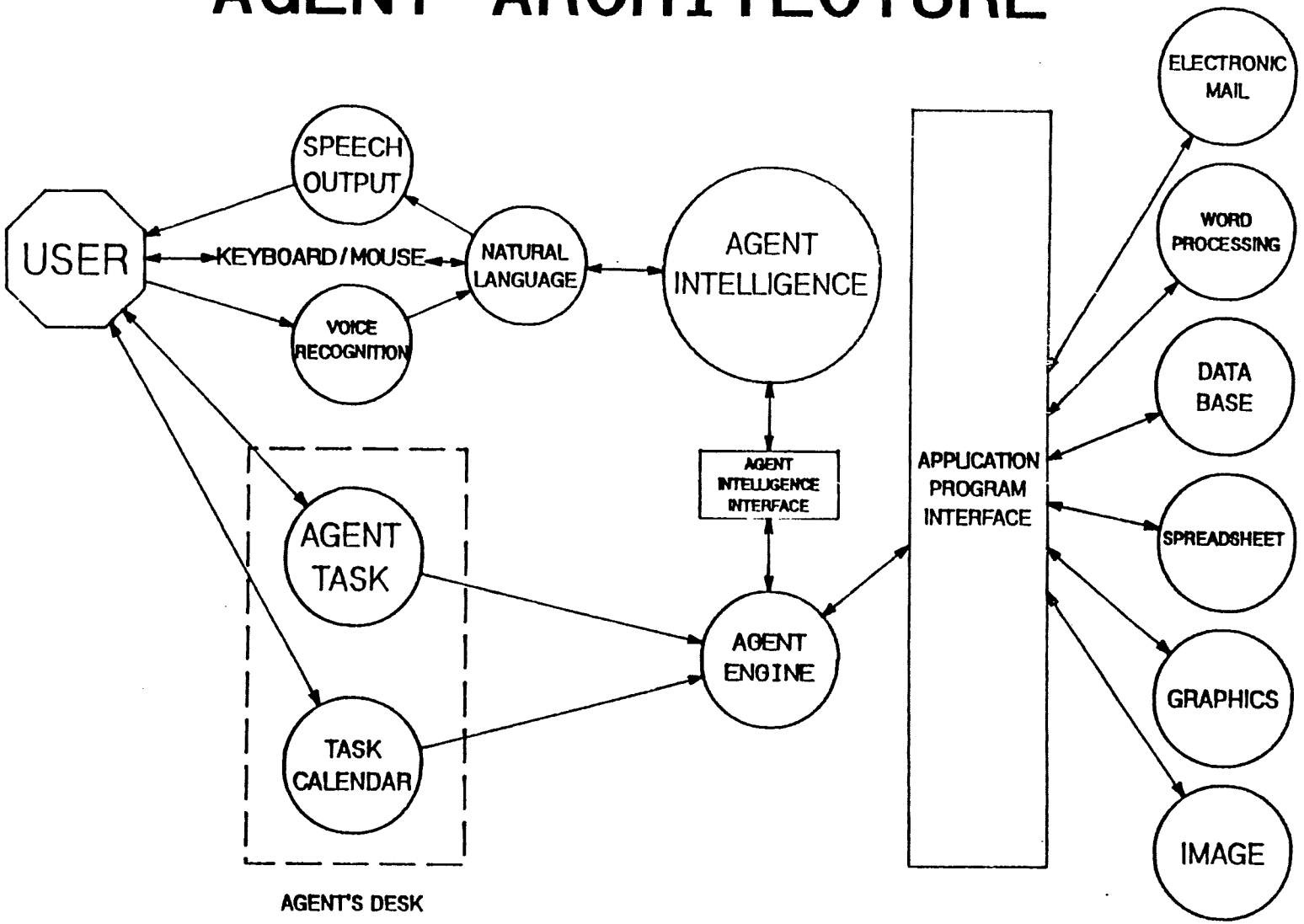
AGENT EVOLUTION

(Toward a Software Paradigm)

- PHASE I – API for Developers
- PHASE II – Agent for CBT
- PHASE III – Procedural Agents
- PHASE IV – Remote & Background Procedural Agents
- PHASE V – Expert Procedural Agents
- PHASE VI – Animated Talking Procedural Agents
- PHASE VII – Intelligent Agents For Workspaces
- PHASE VIII – Intelligent Agents For Applications
- PHASE IX – Cooperative Intelligent Agents With Other Intelligent Agents
- PHASE X – Cooperative Intelligent Agents With Other Users

Building the vision a release at a time.

AGENT ARCHITECTURE



HEWLETT PACKARD

QUESTIONS

* WHAT DOES IT MEAN TO HAVE A 'PRINCIPLED' AGENT ARCHITECTURE ?

* WHERE DO WE GET THE PRINCIPLES FROM ?

* ARE WE MODELLING THE RIGHT THINGS AND DOING JUSTICE TO THEM ?

* WHAT IS THE MOTIVATION BEHIND MUCH OF THE WORK ON AGENCY : SKILL MODELLING OR PSYCHOLOGY ?

Current research of Phil Stenton: Designing Cooperative Interfaces: Tailoring the Channel

Phil Stenton
Hewlett Packard Labs., Bristol BS12 6QG

Abstract

This short paper discusses our work on designing cooperative interfaces to information management systems. We have not taken the common approach to this problem of picking an interesting cooperative response and building a system to generate it. We have analysed real data from experiments and field study transcripts. We have used the resulting dialogue theories to match interface technologies to dialogue requirements. To test our theories we have designed and built a financial information system (GAP). In the process, we have customised two broad coverage NL systems (DataTalker from NLI and the Core Language Engine from the SRI/Alvey program) and integrated one of these within a direct manipulation interface.

1 Introduction

The work on cooperative problem solving has three main foci:

1.1 Cognitive Load Distribution

The first focus stems from the observation that 'Expert Systems' do all the problem solving and the user provides the data, whereas information systems provide all the data and the user does all the problem solving (Kidd 1984,85 Allport 1989 Miller 1984, Cohen & Levesque 1987, Allen & Perrault 1980). The work from this camp is aimed at redistributing the problem solving to somewhere in the middle.

1.2 Dialogue Control

The second focus is similar to the first but the emphasis is on control of the interaction. The two ends of the cooperativity spectrum are system controlled and user controlled. The key to cooperativity is the distribution of dialogue initiative (Reichman 1985 Frohlich et al 1987, Gilbert et al 1987, Whittaker & Stenton 1988 Matthews 1985). Again the goal is to build systems which are positioned somewhere in the middle, customised to the needs of the problem solving task.

1.3 Channel Tailoring

The final focus centres on the design of the communication channel to make the information exchange as easy as possible. Cooperative behaviour is thus supported through the surface characteristics of the interface (Sneiderman 1986, Miller 1987, Williams 1984, Whittaker &Stenton 1989 Gross 1977 Gross & Sidner 1987, Walker 1989). In this group we include the NL discourse work.

2 A List of Question-Answer pairs is not enough

Common to research on each focus is a requirement for extended dialogue data. That is, data from dialogues which focus on problem solving tasks and extend beyond single question-answer pairs. In the process of our research we have worked on each of the three foci and collected dialogue data along the way. Our main (and current) contributions have been in the last group, Channel Tailoring. Our goal has been to design interfaces to information systems which support the kinds of extended dialogue users WOULD LIKE TO HAVE. Our target user group is business executives and our task focus is marketing and sales information management.

3 Intuition and interviews provide only half the story

Most of the data which has encouraged research on cooperative interfaces owes more to interviews and intuition than empirical observation. Exceptions to this include Grosz (1977), Guindon(1986), Kidd(1984), Jarke et al (1985), Dahlback & Jonsson (1989), Whittaker & Stenton (1988, 1989).

Collecting dialogue data is not easy. Collecting it from busy business professionals who have real time problems to solve is almost impossible.

Interviews are sufficient for getting at the 'Strategic' requirements. They tell you WHAT information is required by a population of users but not HOW the interface should facilitate its retrieval. For some work on Cognitive Load Distribution it may be sufficient to know that users want to ask the system certain questions. The inferencing capabilities of the system can then be determined from the expected answers (e.g. Kidd & Allport 1989).

To build mixed initiative dialogue systems or cooperative communication channels requires 'Process' data. That is, data about HOW users would like to interact with a cooperative system. It extends beyond simple question- cooperative answer pairs of the sort developed by Motro 1986, Mays 1980 Kaplan 1982 and others. Process data provides information about the routes that dialogues might take and the information that is required to facilitate a user following those routes. For example, during a dialogue a context may be created of all the objects mentioned. Users may want to refer to those object without having to describe them in full every time. Thus a dialogue model as described by Schuster (1989) would facilitate this behaviour. A more sophisticated model might seek to order these items to reflect the attentional state of the dialogue (Grosz & Sidner 1986). Another example of the use of process information might be in the choice of interface modes to suppport the dialogue. The requirements for deictical gestures and/or detailed NL descriptions differ between and within dialogue routes. The spatial resolution of deictical reference may also differ (Wahlster 1988). Interviews rarely reveal these requirements.

4 Tailored integration is worth 1000 inference rules

The primary focus our work has been the development of dialogue theories for business information systems. We have collected both Strategic and Process data and identified the benefits of different interface modes for supporting the resulting dialogue requirements.

The strategic data we acquired through interviews, in the usual way. The process data was obtained by running Wizard of Oz (WOZ) studies and the analysis of transcripts from real-time sessions with a Natural Language system (Intellect). The latter NL transcripts were supplied by a customer and were generated by managers during the course of their business.

Our WOZ analysis is reported in Whittaker & Stenton (1989) and the resulting dialogue theory in Stenton et al (1988). The advantage of WOZ experiments over protocol analysis is the freedom from the limitations of existing interface technology. The only limitation was that subjects had to type their input. It is true, this did not allow pointing, but it became clear where subjects would have benefitted from this facility. From these experiments we were able to identify not only the limits of NL technology, but also its appropriateness. That is where its use was appropriate and where Direct Manipulation was better and how the two could be combined to facilitate the transfer of information.

¿From the customer transcript analysis we were able to identify the benefits that users of an existing NL system were getting over an available alternative (menu based system). We were also able to address the problems of customising a large NL system (Intellect was replaced by DataTalker). We identified the mismatches between Intellect's capabilities and the things users tried to do with it. User's were found to require specific features from the three categories of Selectivity, Flexibility and Presentation (for a description of these categories see Walker 1990).

5 Mixing metaphors by design

Tailoring the communication channel to support cooperative communication is stretching the currently popular deskstop metaphor beyond its limits. The propogation of intuition afforded by this metaphor is already meeting inconsistencies. These usually appear where the features of electronic storage and manipulation are exploited (e.g copying documents by pointing at them while holding down a combination of keys).

Interfaces which are limited to desktop metaphor are confined to situations where the initiative must always be with the user (what does it mean to have a desktop which takes the initiative on occasions?). The introduction of Natural Language will also be a challenge for the consistency of the desk top metaphor. From our experiments it is clear that Natural Language has an important role for information retrieval systems. If we tailor the interface to such systems to suit the required dialogues we must include Natural Language. Does this mean we have to believe in a talking (or at least listening) desktop ?

The paradigm shift in interface design caused by the use of strong metaphor made computers accessible to a larger population. Stretching the chosen metaphor to make available the work on cooperative interfaces is likely to distort it. As a result the power of the interface to stimulate users' intuitions will be weakened and the paradigm shift will have been short lived.

There are two approaches to tailoring the interface to support cooperative behaviour and preserve the paradigm shift: The first is to change the metaphor to one that is more appropriate (Schon 1982); The other is to mix metaphors in a sympathetic way. It is here that the Agent metaphor has a role to play.

We have been working on the notion that the desktop metaphor is limited but irreplaceable. The solution then is to mix-metaphors by design. That is to find a suitable metophor to facilitate the intuitions of the users as the desktop does for simple paper shuffling tasks. The Agent is the best metaphor around at the moment but its acceptance is hampered by the runaway visionaries who use the notion as a repository for AI magic.

A focus on dialogues as a spin off from the agent metaphor looks promising to us. It emphasises the interaction and not the cognitive load sharing aspects of agents. The agent would have access two classes of knowledge: information it has but the user 'might not', which includes meta-

knowledge about the rest of the world (things not on the desktop); and mutual knowledge about past conversations, commitments (To Do list); and the objects in the current dialogue.

The role of an agent in our model is that of an 'assistant' (Kiss 1987) where the agent is the interface to a distributed environment providing the tools necessary to for the user to perform tasks. The agent metaphor is used here to stimulate users' intuitions about dialogue context and the agent as an assistant who has a limited memory for past dialogues and can 'understand' a subset of Natural Language. The user modelling facility of our agent is limited to objects mentioned (used or created) in the current or previous dialogues or visible on the desktop. We are currently working on a mixed-mode interface to such an agent.

6 Conclusions

Work on cooperative problem solving systems can be described as having three broad foci: Sharing the problem solving through distributed inference; sharing the control of the dialogue through mixed initiative; and facilitating the dialogue by tailoring the surface features of the interface. All three require data to be collected about HOW users would like to communicate with system, though some work on cooperative responses has been carried out in the absence of empirical observation. Our work has identified the benefits of different interface modes and how they can be combined to facilitate the cooperation between business executives and information systems. We have instantiated our dialogue theories in a financial information system (GAP) which is under evaluation. We are currently working on an agent metaphor as the repository for our theories.

7 References

(for references not included here see Stenton 1987):-

Allport D. (1989) A Computational Architecture for cooperative systems, Invited Paper to appear in *Proceedings of KBCS Conference*, Bombay, India

Allport D. & Kidd A.L. (1989) Using knowledge about search spaces to give cooperative responses. *Hewlett-Packard Tech Memo HPL-ISC-TM-89-129* Frohlich D.M., Crossfield L.P. & Gilbert G.N. (1985) Requirements for an intelligent form-filling interface. In P.Johnson and S. Cook (Eds.) *People and computers: designing the interface.* Cambridge Univ. Press

Gilbert G.N., Luff P., Crossfield L.P. & Frohlich D.M. (1987) A mixed initiative interface for expert systems: The Forms helper. *International Journal of Man-Machine Studies*, forthcoming.

Grosz B.J.& Candace L. Sidner. (1986) Attentions, intentions and the structure of discourse. *Computational Linguistics* 12 pp. 175–204

Kiss, G (1987)Why bother about goals ? *2nd Intelligent Interfaces Meeting (IISIG) London pp 66-77 (Invited talk)*

Matthias Jarke, Jon A. Turner, Edward A. Stohr, Yannis Vassiliou, Norman H. White, and Ken Michielsen. A field evaluation of natural language for data retrieval. IEEE Transactions on Software Engineering, SE-11, No.1:97–113,

Schon D. (1982) The Reflective Practitioner MIT Press

Schuster E. (1989) Establishing a relationship between discourse models and user models Computational Linguistics vol 14 no 3 pp 82-85

Shneiderman B. Software Psychology. Winthrop Publishers Inc.

Stenton S.P. (1987) Dialogue management for cooperative knowledge-based systems Knowledge Engineering Review vol 2 no 2 pp 99-121

Stenton S.P (1988) Supporting set manipulation dialogues for information retrieval Hewlett-Packard Tech. Report HPL-ISC-TM-89-02

Walker M.A. (1990) Natural Language in a desktop environment To appear in the Proceedings of HCI International Boston Sept,1989

Whittaker S.J. & Stenton S.P. (1988) Cues and control in expert-client dialogues Proceedings of the 26th ACL pp 123-130

Whittaker S.J. & Stenton S.P. (1989) User studies and the design of Natural Language systems Proceedings of the 4th EACL pp 116-122

Current research of Ann Blandford:

Ann Blandford
IET, Open University

I am currently working on the development of a model of tutorial dialogue based on aspects of agent theory. This work is motivated by an interest in deriving tutorial interactions from a deeper representation of the goals and beliefs of the participants than has hitherto been the case in ITS research.

The domain in which this work is based is a topic within decision analysis (Multi-Attribute Utility Theory). In this context, information is not 'certain', and is more appropriately dealt with as 'justified beliefs'. The computer tutor seeks to engage the student in an 'animating' dialogue, discussing aspects of both the current decision problem (e.g. which factors should be considered in reaching a decision, and why) and general decision making strategies (e.g. what is an appropriate action to take next, and why).

The interaction is not natural language, being based on a formal representation of the locutionary force of an utterance.

A skeletal 'action cycle' for a (computer) dialogue participant based on its values, beliefs, wants and commitments has been implemented. Its values include both local values (valuing having done something) and longer term values (such as valuing keeping the student motivated). Its beliefs include, for example, its beliefs about:

- the user's beliefs about the problem (including the grounds on which it holds that belief - e.g. that the user proposed a factor, or that the user implicitly accepted a suggestion from the system),
- the factors they have agreed to include so far, whether or not both parties agree about their importance. These are referred to as 'mutual working beliefs'.
- the user's performance, in terms of aspects such as the 'quality' of suggestions made by the user.
- the user's wants, as expressed through the dialogue.
- various aspects of the problem.
- what actions can achieve progress towards valued states.

The 'action cycle' consists of identifying reasonable acts to do in the current context, deciding which to do (based on the system's values and beliefs), and doing it. Non-primitive actions have sub-parts, the doing of which involves making further decisions about how to achieve them. In practice, there are only three non-trivial decision points: deciding whether or not to listen to the user, deciding whether of not to respond to the user, and deciding what to say to best satisfy the system's values. (This includes valuing "keeping the user happy" - i.e. taking into account the system's beliefs about the user's values/goals.)

Work over the next few months will involve integration of other essential components with the action cycle, with a view to empirical testing of a 'rational tutorial dialogue agent' towards the end of this year. These components include:

- a mechanism for identifying reasonable responses (including heuristics to limit the number of such responses)

- a decision mechanism for deciding which of the possible actions or responses is preferred

- plausible reasoning mechanisms (to assess the user's proposals and construct lines of argument)

- ability to interpret the meaning (or illocutionary force) of the user's utterance based on the system's expectation (e.g. if the system has requested a justification then it has an expectation of the user giving one) and the locutionary force of the utterance

- belief structures to encode information about the current problem and decision making strategies

This is of necessity a brief outline of my current work. Details will undoubtedly change as implementation progresses!