

# CoSy Meeting Freiburg 1-3 March 1005

Aaron Sloman

<http://www.cs.bham.ac.uk/~axs/>

School of Computer Science

The University of Birmingham

---

## ARCHITECTURES FOR COSY ROBOTS

### SUMMARY

Objectives and Overview

The short term: Mr Chips I and II

The long term: PlayMate and Explorer (and Philosopher?)

The longer term: Combining PlayMate, Explorer and Philosopher

**The need to understand requirements and design options**

**Scenario-based task analysis**

**Some choices to be made**

Please send me suggestions for improvement.

# Architectures and Science

---

- Engineers typically start from a more or less well-defined practical problem or goal and attempt to find a solution.
- If the solution works, is reliable, robust, maintainable, and not too costly, everyone concerned is satisfied.
- However this does not imply that any knowledge has been gained, apart from knowledge of how to solve that problem and very similar problems.
- Scientific understanding requires something more principled: a different level of generality, which advances knowledge in a way that explains something more general about mappings between problems and solutions.
  - This obviously does not rule out producing working systems or solving practical problems.
  - But such achievements are not the *primary* goal – they are justified insofar as they help to drive the science and test it.
  - According to Colette Maloney (June 2003) applications serve to
    - \* provide research questions
    - \* demonstrate impact of conceptual/technical innovation
- We need to have a clear vision of what we are trying to achieve and how to get there: not just a bunch of ‘look ma no hands’ demonstrations putting together things we already know how to do.

# TWO TRAPS TO AVOID

---

- In principle the project could fall into the trap of producing a collection of vacuous high level theories: but we are committed to going beyond that by firmly adopting a scenario-driven approach, where each scenario requires a working system which can be demonstrated on a range of tasks of sufficient generality to test the advances in knowledge and demonstrate the novelty of the ideas.
- That requires the demonstrations to include performances that nobody now knows how to achieve.
- We could fall into the alternative trap of merely exercising ingenuity to produce some demonstrations – what John McCarthy called the ‘Look Ma, no hands’ approach to AI, and Carl Hewitt summarised as ‘A hairy kludge a month’
- One way to avoid this is to have
  - a well-defined trajectory through a succession of increasingly difficult tasks, each of which provides extra demands on theories, formalisms, mechanisms, and architectures
  - related to a general conceptual framework for describing and evaluating architectures
  - ideally also related to explaining aspects of natural intelligence.

# What we said we would do

---

## Extract from the project summary (Colette Maloney's vision?)

We start from the assumption that the visionary FP6 objective

**“To construct physically instantiated ... systems that can perceive, understand ... and interact with their environment, and evolve in order to achieve human-like performance in activities requiring context-(situation and task) specific knowledge”**

is far beyond the current state of the art and will remain so for many years.

However we have devised a set of intermediate targets inspired by that vision.

Achieving these targets will also provide a launch pad for further work on the long term vision.

In particular we aim to advance the science of cognitive systems through a multi-disciplinary investigation of requirements, design options and trade-offs for human-like, autonomous, integrated, physical (e.g., robot) systems, including requirements for architectures, for forms of representation, for perceptual mechanisms, for learning, planning, reasoning, motivation, action, and communication.

**We did not promise any particular performance: the main aim is new scientific understanding.**

# Why it is hard

---

Nobody has done this, or even come close (as far as I know).

- **The scientific and technical problems are very hard, e.g.:**
  - How do we see 3-D surface structure, and positive and negative affordances?
  - What should be innate, and what learnt or developed, and how?
  - How should the robot represent what does not exist but *might*
  - How should the robot be able to think about and explain its own thinking, or refer to someone else's thinking – using meta-semantic competence?
- It is not obvious what the state of the art is in all relevant areas — though we all know different aspects of the state of the art.
- Minor problem: research staff starting late, and equipment not yet available.
- Not so minor problem: We have not worked together previously.
- Different partners have different constraints and commitments.

**THE HARDEST THING IS TO UNDERSTAND WHAT ALL THE PROBLEMS ARE: ANALYSIS OF SCENARIOS WILL HELP US IDENTIFY THEM, AND EVALUATE SOLUTIONS.**

Note: many AI predictions failed because people mistakenly thought they knew what problems they were addressing.

# PARTIAL SOLUTION

---

**People working on Playmate started off with an easier task**

- **in order to get to know one another**
- **in order to gain some experience of integrating different sorts of functionality**
- **in order to explore some of the available tools and techniques including**
  - **tools for connecting machines and processes (Nick will talk about this)**
  - **speech understanding and generation tools**
  - **parsing and sentence generating tools**
  - **vision tools**
  - **planning techniques**
  - **ontologies required by the robot**

**But we should be prepared to throw away the results.**

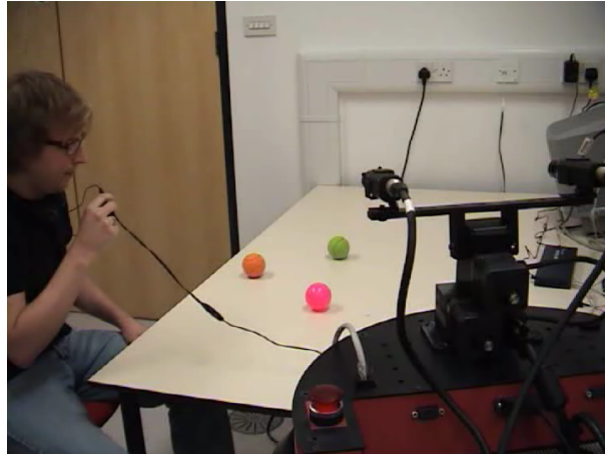
# How we have worked

---

- Initially we set up a very simple demo, Mr Chips I, involving only BHAM and DFKI

**This produced the video (6th Jan)**

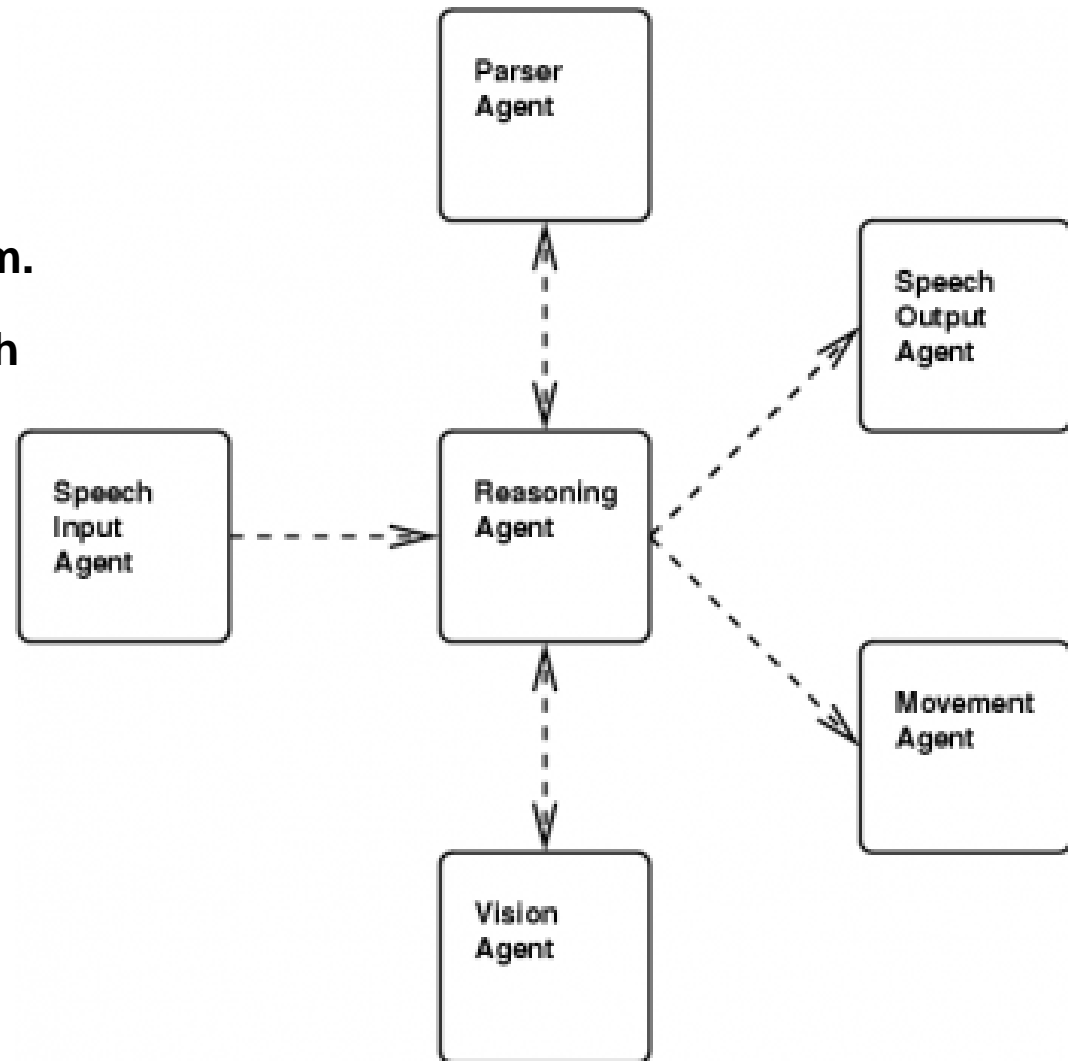
**Has everyone seen it?**



- Later, more groups met in January in Saarbrücken to plan **Mr Chips II**, adding more sophistication to all the components.
- **Both Mr Chips I and Mr Chips II are merely transitional, i.e. steps towards understanding requirements and possibilities for PlayMate, and learning about available techniques and their limitations.**
- In parallel with the design and implementation work we have done a lot of reading and discussion of much broader issues, e.g. in biology and psychology.

# Mr CHIPS I

- A couple of weeks' work by people at DFKI and Birmingham.
- Provided experience using both Marie and OAA to combine different sorts of software.
- Movie impresses people, but there is no new science.
- Still very useful for people involved to get started.



# Mr CHIPS II

---

## A larger group discussed this in Saarbrücken

(BHAM, DFKI, CNRS, TUD, ALU-FR)

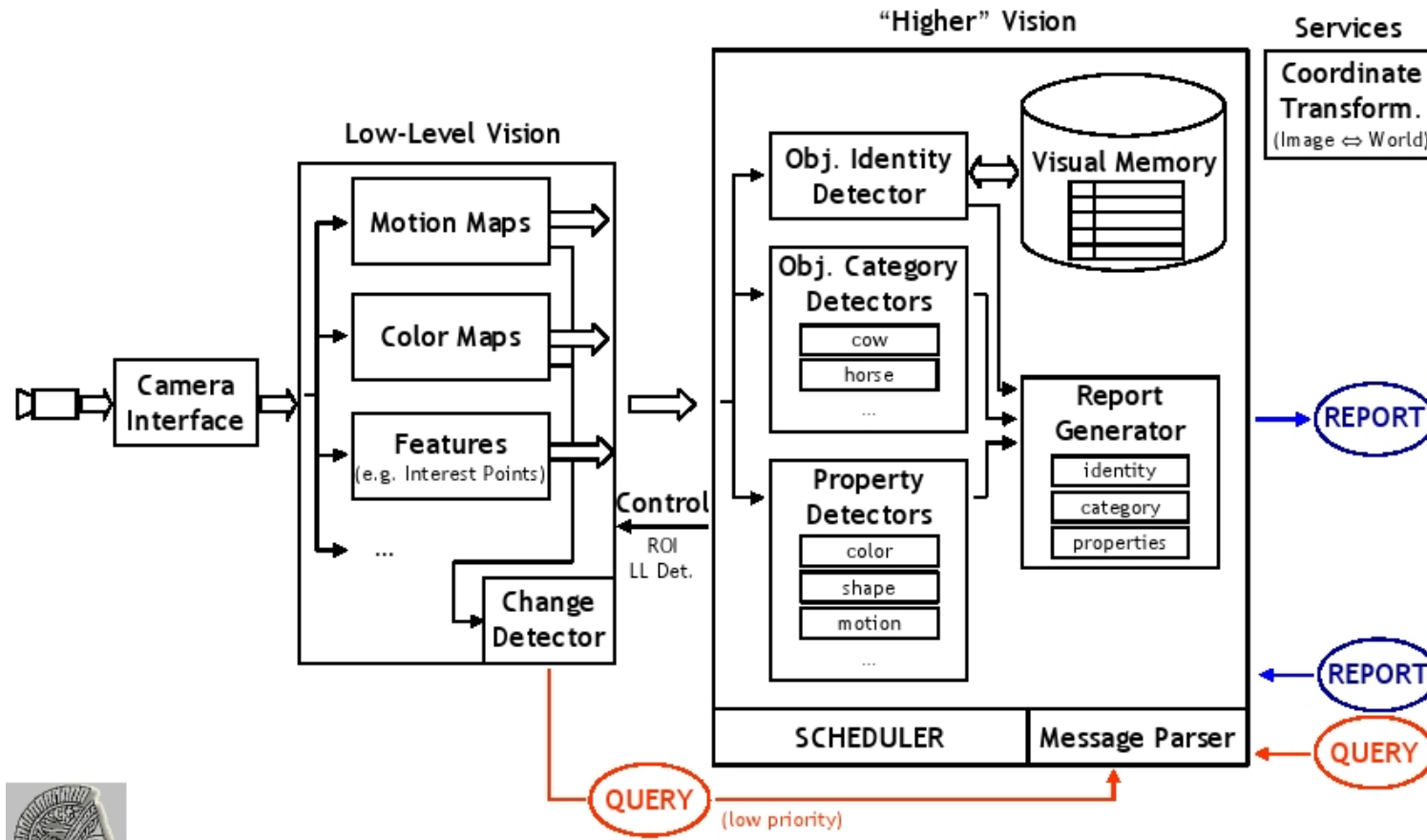
- Adding more sophisticated vision
- Extending the dialogue
  - More syntactic forms
  - Dealing with ambiguity
  - Planning what to say (e.g. how to ask a question to resolve ambiguity)
  - Describing actions as well as objects and relationships
- Adding more complex actions
  - E.g. while we are still waiting for the arm, make the head move.

**Jeremy has produced a quite long and detailed specification for Mr Chips II.**

**(Ideally we need an 'Explorer status report' here.)**

# Vision system sub-architecture for Mr Chips II

Plans for vision system developed at DFKI meeting: TUD vision for Chips II.



Many unknowns still: what do we mean by shape – could be trivial or impossibly difficult? How much concurrent processing at different levels? Should there be variable resolution receptive fields? How much top-down influence? (more later)

Jeremy and GJ will say more about other plans for Mr Chips II

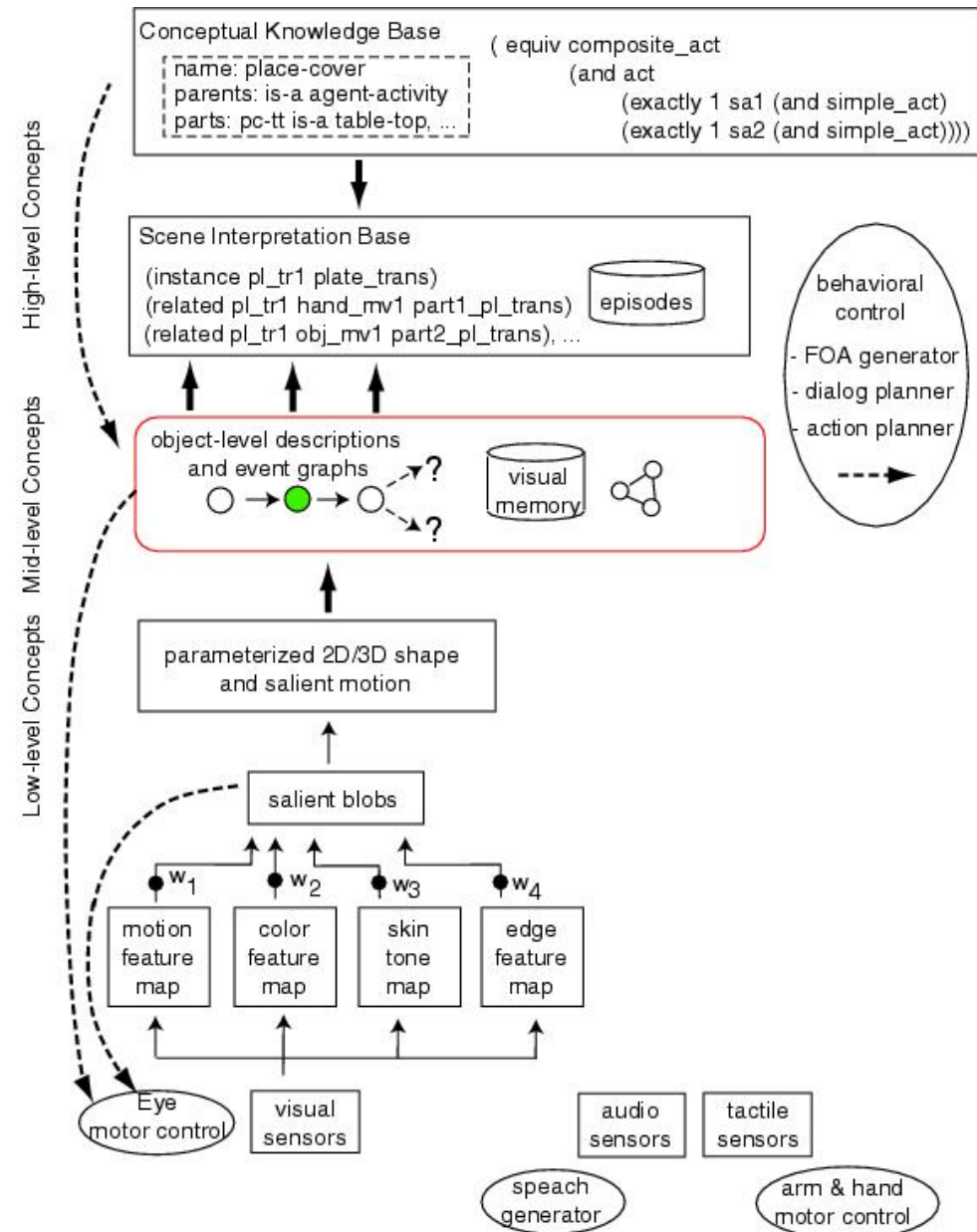
# Another visual architecture

## Somboon's work:

This is concerned with the perception of motions, some of which are intentional actions.

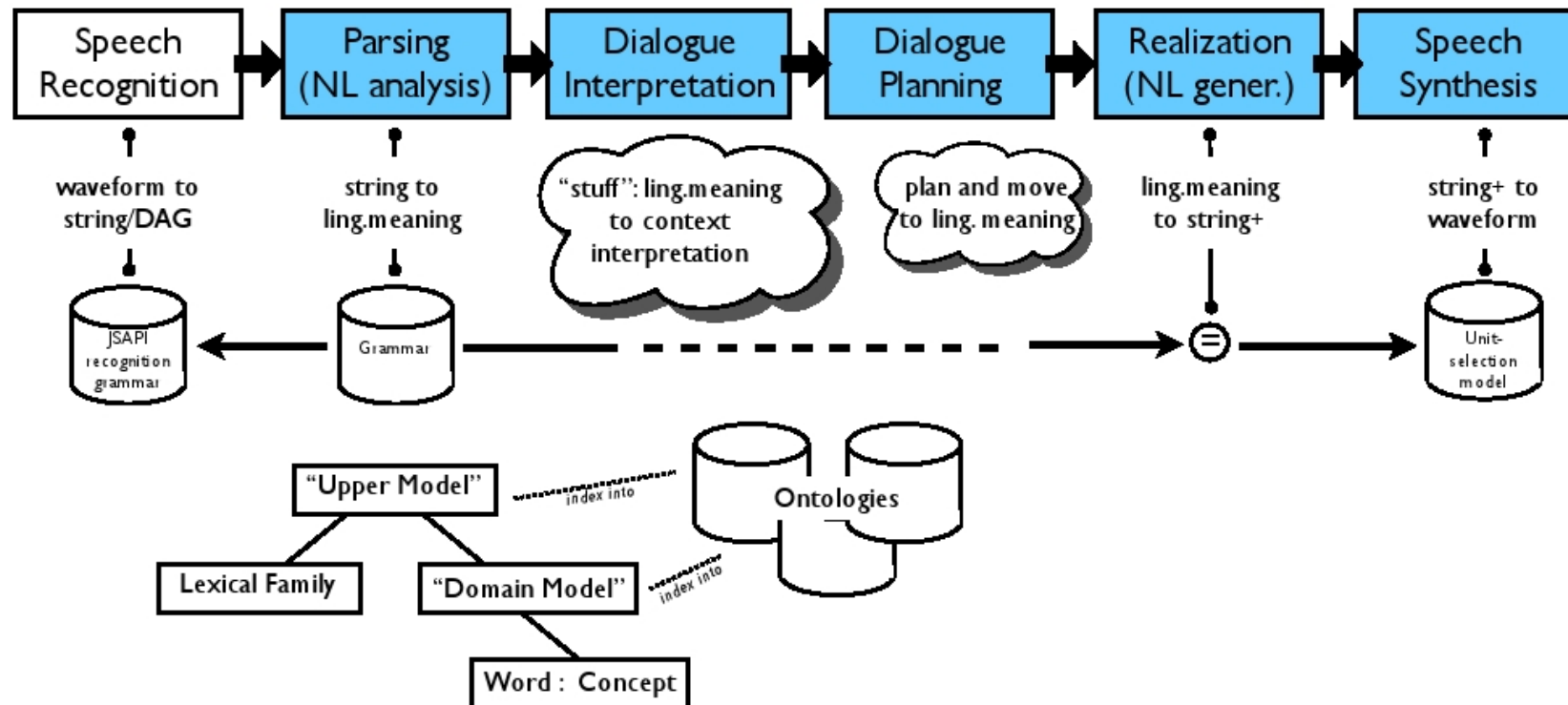
The ideas can be combined with those of the TUD architecture or pursued in parallel producing different versions of Mr Chips 2, demonstrating different capabilities.

But in the end we need something more principled, more general, more integrated.



# Possible Linguistic component for Mr Chips II

The previous architectures are being discussed in the context of a system that is primarily a linguistic agent: it will talk about what it sees and answer questions about what it sees.



Here is a possible linguistic sub-architecture (from a document by GJ). It is not clear how this should interface with an architecture that might have developed independently for a complete pre-linguistic agent. That is something to be investigated, within the framework of a general theory about architectural components, and types of designs.

# Towards a general theory

---

In parallel with the design and implementation work we have done a lot of reading and discussion of much broader issues, e.g. in biology and psychology.

In particular, with biologist Jackie Chappell, who led the work on the New Caledonian crows (Betty and Abel) in Oxford, and is now in Birmingham, we are attempting to analyse the altricial-precocial spectrum in a way that is equally applicable to animals and to robots.

An incomplete draft paper is here

<http://www.cs.bham.ac.uk/research/cogaff/altricial-precocial.pdf>

If our robots are to be near the altricial end, able to make new discoveries, learn new concepts, understand new goals and tasks, and not merely learn by adjusting parameters, then we'll have to investigate appropriate sorts of mechanisms, architectures and forms of representation.

**For example it seems that even pre-linguistic altricial animals seem to need something like syntactic competence for internal representations in order to be able to create new combinations from old information units.**

# Beyond Mr Chips

---

**Mr Chips II is centred round linguistic interaction: requirements for language processing dominate the architecture.**

**For example, there is emphasis on objects, their properties and relations, since those are things that we naturally talk about, and most of the work on planning in CHIPS II is concerned with planning what to say or to ask.**

**So in parallel with work on Mr Chips II a few people at BHAM and CNRS (also talking to non-CoSy colleagues at BHAM) have been looking ahead to a task after Mr Chips involving a visually and behaviourally more competent **pre-linguistic version of PlayMate** (Mr Chimp???), inspired by crows, chimpanzees, and children aged about 1 to 3 years, where vision guides and suggests non-verbal behaviour.**

**(Show yoghurt movie?)**

**This has raised important questions about the nature-nurture issue, leading to investigations with Jackie Chappell (Biosciences) on the precocial-altricial spectrum in nature and how it might apply to robots.**

Jackie previously worked on Betty, the hook-making crow.

IJCAI paper and poster submitted: proposing an outline 'high-level' theory of the mixture of precocial and altricial skills required for a PlayMate-like robot.

**There is a lot more work to be done — grant proposal under discussion.**

# Precocial/Altricial

---

- **Precocial**

Some animals are born or hatched highly competent: deer, chickens, etc.

- **Altricial**

Some animals are born/hatched underdeveloped and highly incompetent, but adult forms can do things precocial species cannot, e.g. hunting mammals, nest-building birds, primates, humans.

- **Actually it's not so simple: even altricial species have some precocial skills or tendencies, e.g. sucking, stimulating parents to feed, and some 'delayed' precocial skills, e.g. sexual maturation in humans.**

- So what can be or needs to be learnt is a very subtle issue.

- Architectures and competences may be pre-formed in precocial species, but slightly adaptable, e.g. by reinforcement learning.

**contrast learning a language, or learning to program computers**

- **Altricial species may be using sophisticated architecture-growing mechanisms doing far more than varying weights (etc.) when they look incompetent.**

Perhaps collecting chunks of information about affordances provided by the environment and their bodies — initially stored then later recombined and used.

Perhaps building sub-architectures tailored to the environment.

# Limitations of current theories of learning

---

- The majority of current AI work on learning on learning seems to be statistics-based associative learning.
- This may be part of the explanation of the ‘minor’ advances and adjustments in skills of precocial animals.
- But it fails to explain the advances made during a life time by animals with altricial capabilities (e.g. humans, chimps, crows), such as
  - learning a new form of representation (internal or external)
  - learning a new or an extended ontology
  - learning creative new ways of combinin old skills
  - learning new ways of thinking.

**We need to design scenarios and architectures addressing the precocial-altricial spectrum.**

**This requires collaboration with developmental psychologists and biologists to ensure that we have sufficiently deep and broad understanding of what needs to be explained.**

## **NOTE:**

**We should not assume they already have good theories that we merely implement: most don't know how to design working systems of the sorts they are trying to explain.**

# Precocial species refute naive 'symbol-grounding' theories

A highly competent calf has not had time to derive concepts from experiences: semantic capabilities must arise from genetically determined structures.

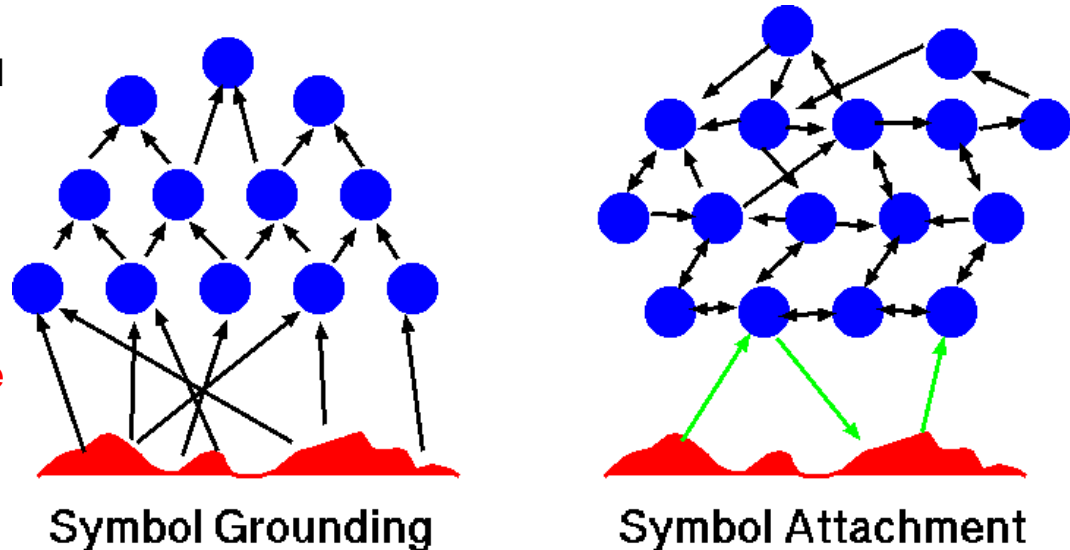
The idea of **symbol-grounding** (concept empiricism) is that all meaning is derived from ('flows upward from') sensory processes (experiences)

The idea of **symbol attachment** is that a great deal of meaning comes from the structure of a theory and internal constraints, limiting possible models (**we need to generalise Tarskian semantics**).

Bridging rules (links with perception and action), attach the structure to the environment and thereby help to reduce or remove residual indeterminacy of reference and make the theory applicable. (Standard 20th Century philosophy of science.)

It's more complex than that: we have many mutually supporting theories. Moreover, because a theory uses a reasoning system, new attachments (e.g. types of experiment or application) can be derived from old ones: the source of explanatory and predictive power of theories.

See <http://www.cs.bham.ac.uk/research/cogaff/talks/#grounding>



# Why a child, or a chimp, needs deep theories

Perceivers like us need some understanding of how shapes, colours, weights and other properties of things change as various things are done in the world, and how our experience of them should change while we move or act and the context (e.g. lighting) changes, though the objects persist unchanged.

Kant: we have to assume there is a reality whose properties persist independently of our changing experiences of them – as the hidden properties of atoms, sub-atomic particles, force-fields, genes, niches, evolutionary pressures, economic and social forces, etc. persist independently of our changing experiences of them.

Something like this seems to be a requirement for understanding some of the most mundane aspects of our world and some of the deepest.

It is often assumed that all such knowledge of the environment must come from experience, by various kinds of induction or abduction, or statistical inference.

**That assumption could be false if evolution produces structures that constrain what the individuals of certain species are capable of learning, so that not all details of the theory of the environment have to be learnt empirically (as in precocial species), or new theories, not derived from experience, are invented.**

Note: whether what is learnt or innately determined is **implicit** (only embodied in the design of the system) or **explicit** (represented in some formalism that the system can manipulate, e.g. in making inferences) is a separate issue.

In microbes and insects it is probably all implicit. Some mammals and birds may be different.

# Naive innatism almost fits precocial species

If genes can produce at birth appropriate structures (theories), and mechanisms for using them, then, insofar as those structures allow some things in the environment as models and others not, the structures may determine much of what the concepts used connote.

So innately determined structures (used via links to sensors and motors) may provide all or almost all the meaning required by members of a precocial species. Indeterminism and ambiguity of meaning may be removed or reduced by (implicit) rules linking innate structures to perceptual and action mechanisms.

These help to pin down the interpretation of the theory so as to make it refer to the immediate environment, rather than some isomorphic 'twin earth'.

- This works dramatically for precocial species, e.g. deer that can run with the herd within minutes of being born.
- They don't have time to develop all the conceptual apparatus on the basis of their individual experience.
- Innate structures and manipulation mechanisms can enable a novel configuration of terrain (e.g. a rock ahead) to be understood in such a way as to produce appropriate action – e.g. making a detour or climbing over.

**What about altricial species?**

# **Beyond naive innatism: altricial species**

---

For altricial species – e.g. humans, primates, hunting mammals, nest-building and hunting birds, the genes do not provide all the information used in adult life. (Why not?)

Instead there seems to be a genetically produced ‘bootstrapping program’ that allows the structures to be built during early interaction with the environment, even while the brain is still growing. This may explain why certain kinds of learning have to occur at critical ages (approximately).

**That means the later stages are a product of both nature and nurture, and it is possible for nature to impose strong constraints on what nurture can generate: millions of years of evolution should not be wasted.**

The implication of this would be that the theoretical structure determining the meanings used by an adult of such a species could have many abstract (high level) features determined genetically and common to all members of the species, whereas concrete details are determined by the individual’s environment (and culture).

(Compare Chomsky and others on language universals.)

(Note different reasons why genes leave so much unspecified for altricial species.)

# CONJECTURE

---

- One of the most powerful mechanisms supporting adult sophistication and variability in altricial species is the ability to acquire information **chunks** which can be combined into more complex re-usable chunks.
- If this is recursive it is particularly powerful.
- Transformations of such complex chunks provide mechanisms of inference, planning and interpretation.
- These are essentially 'symbolic' capabilities of kinds required for deliberative mechanisms.

Everything a deliberative system can do can be done by a purely reactive system: provided that all conditions and all actions have been anticipated.

But I assume we wish not to go down that route.

**What sorts of deliberative mechanisms will PlayMate and Explorer need?**

# Some intermediate cases

---

Terminology is often confused and inconsistent in discussions of these matters.

- A system able to construct and manipulate explicit meaningful structures of unbounded complexity, to construct plans, predictions, hypotheses, questions, explanations, using compositional semantics, is often described as having **deliberative** mechanisms, because it is capable of hypothetical reasoning.
- But there are many intermediate cases in the evolutionary history, to which different people apply different labels.
- E.g. some people use the word 'deliberative' to describe a reactive system in which two or more action tendencies (e.g. flee or fight) can be simultaneously activated, with a mechanism to make sure only one of them wins out. I prefer to call these **proto-deliberative** mechanisms, if they don't include the full range of capabilities listed on previous slides.

We need to investigate many intermediate cases including:

- **more or less explicit representations of goals or needs**
- **more or less explicit representations of internal context**
- **more or less creativity in the organism's mechanisms for combining information of different kinds.**

# FULLY DELIBERATIVE SYSTEMS

---

We can define 'fully deliberative' systems as having at least the following capabilities:

- The ability to represent what does not yet exist, or has not been perceived.
- The ability to use representations of varying structure
  - using compositional semantics supporting novelty, creativity, etc.
- The ability to use representations of potentially unbounded complexity  
(Compare fixed size vector representations)
- The ability to build representations of alternative possibilities, compare them, select one.

Recently researchers have started adding reflective and meta-management capabilities, i.e.

the ability to be aware of internal processes, to categorise, evaluate, modify internal processes (reflection/meta-management)

These may use a mixture of reactive and deliberative mechanisms.

# Steps towards PlayMate

The CogAff schema in our workplan provides a first-draft framework for many **concurrently operating** components of an architecture

including allowing for 'multi-window' perception and action, in which processes at different levels of abstraction occur in parallel, with optional reactive 'alarm' mechanisms.

How many of these various kinds of components will be needed in PlayMate?

Will we need reactive mechanisms?

- Low level vision will be data-driven
- Low level motor control
- Simple alarms triggered by either proprioceptive feedback or new input (e.g. something moves or someone talks while PM is thinking or acting)

What sorts of perceptual processes, at what levels of abstraction will be required?

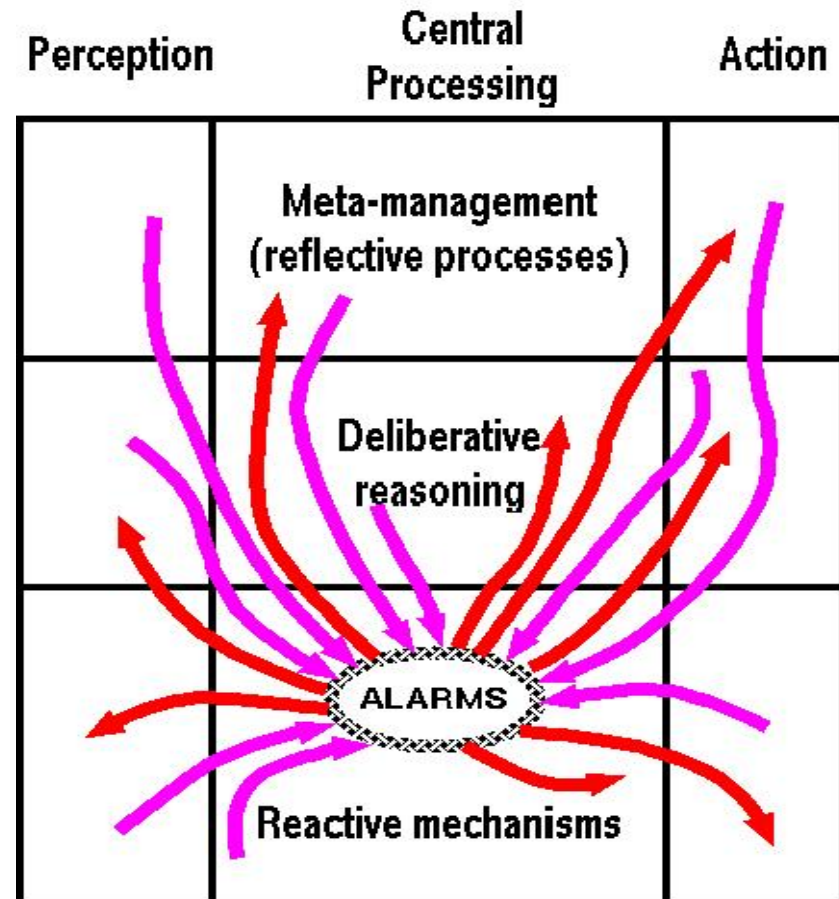
What sorts of actions will be possible?

What deliberative processes?

What meta-semantic processes?

What needs to be innate, and what learnt or bootstrapped?

We need to consider *many* possible scenario-fragments to drive our thinking

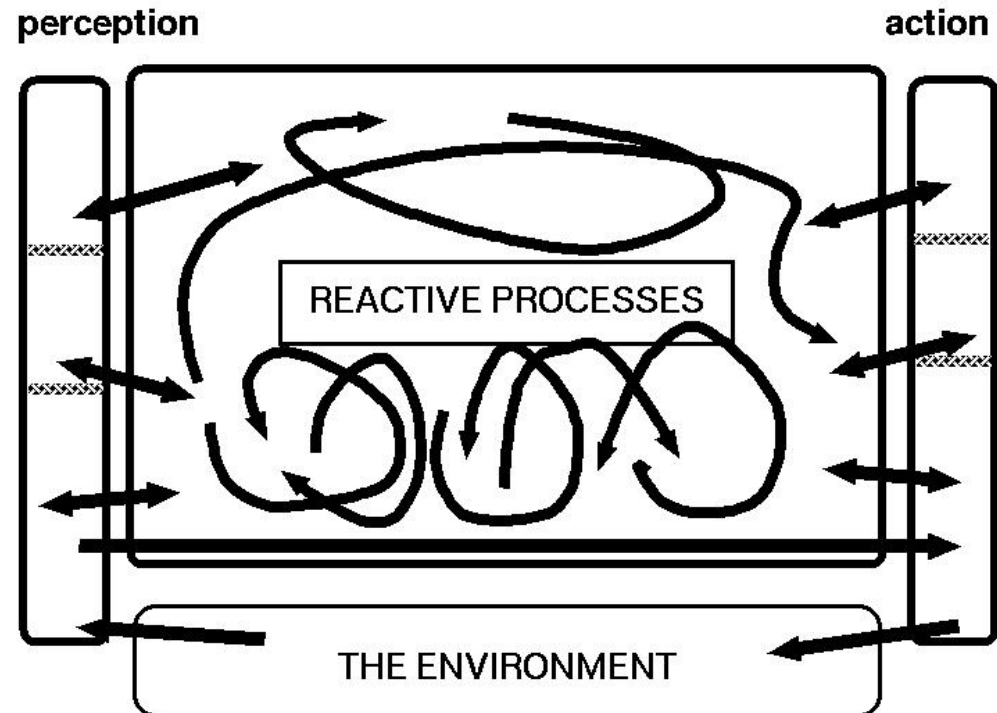


# A simple (insect-like) architecture

A reactive system does not construct complex descriptions of possible futures evaluate them and then choose one.

It simply reacts  
(internally or externally).

An adaptive system with reactive mechanisms can be a very successful biological machine. Some purely reactive species also have a social architecture, e.g. ants, termites, and other insects.



These are **precocial** species: large amounts of genetically determined information, with minor environmentally driven adaptations.

# An 'Omega' architecture uses a subset of the possible mechanisms and routes allowed by the CogAff Schema

A popular design, which turns up with many names, and many different diagrammatic presentations: a special case of the CogAff schema.

Compare the greek Capital Omega letter  $\Omega$ .

This is just a pipeline, with "peephole" perception and action, as opposed to "multi-window" perception and action.

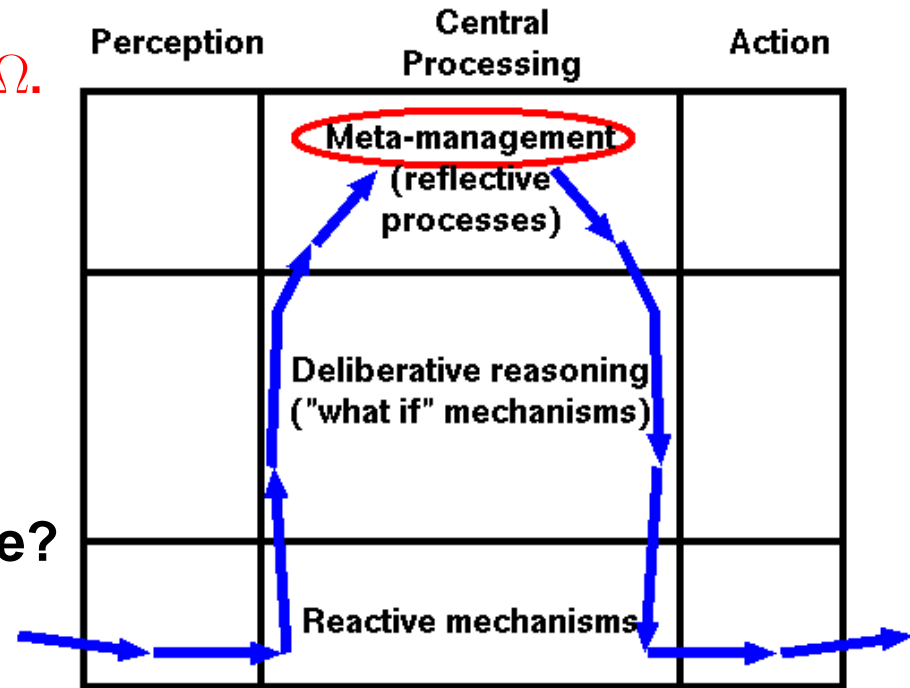
E.g. Norman, Cooper and Shallice: Contention scheduling; and Albus 1981.

Some authors propose a "will" at the top of the omega.

Why will this not work for Playmate?

Various reasons including:

- Need for multi-level perceptual capabilities
- Need for multi-level action capabilities
- Need for various kinds of deliberative processes (including linguistic) processes to occur in parallel with other things, including perceiving and acting
- (Eventually) need for internal self-understanding (meta-semantic competence and meta-management)



# Beyond Mr Chips II

---

- Work on Mr Chips II will continue in order to extend various skills in the project team and increase mutual understanding.
- However, it was arbitrarily chosen to support a particular kind of scenario that integrates most capabilities studied by the partners, apart from learning.
- In parallel with that, analysis and design work is in progress for a pre-linguistic PlayMate, able to perceive and manipulate objects without necessarily being able to talk about them.

Work has started on developing an ontology suitable for a pre-linguistic agent and on specifying requirements for visual perception of shape and affordances, as more basic than object identification and recognition.

This may include some very primitive social interactions.

- Additional work, towards the final PlayMate will consider ways in which pre-linguistic mechanisms can support the development of a collection of linguistic competences and much richer social interactions.
- All our work should take into account the possibility of
  - integration with the Explorer robot
  - extension to the Philosopher scenarios

**The design and analysis will be based on a very large number of mini-scenarios requiring various mechanisms, forms of representation, types of knowledge, types of learning, and architectural features. These scenarios will be partially ordered on the basis of dependences and growing subsets will determine intermediate targets.**

# Scenario-Driven Architecture Design (I)

---

## Examples of pre-linguistic 'micro-scenarios' for PlayMate

- Touching seen objects
- Picking up and putting down objects
- Picking up rotating and putting down
- Stacking and unstacking objects
- Touching a pointed at object
- Picking up a pointed at object
- Putting an object down at a pointed at location
- Stroking a surface
- Moving a finger (or held object) along a groove, or along a curved line.
- Putting a long object into a groove, or an end into a depression
- Pushing an object to a pointed at place
- Pointing at spaces where a certain object could fit
  - **With or without rotation being required**
- Touching an object being moved by someone else
- Tracking (with eye and finger) a moving object
- Pointing at likely place a moving object will hit the table edge, etc.
- Coping with linear, circular, spiral motions, etc., fixed or variable speeds, etc.
- Reacting to an object moving unexpectedly while a task is in progress.
  - E.g. finger, object being moved, target object, etc.
- etc.

**Micro-scenarios also needed for Explorer.**

# Scenario-Driven Architecture Design (II)

---

## Examples of linguistic scenarios for PlayMate

### Examples in Mr Chips II

#### More examples here:

<http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html>

#### Examples needed to drive meta-semantic and meta-management capabilities.

- Why did Nick point at the ball?
- Can Nick see this bit of the table?
- What can't I see?
- Which way do I need to move to see more of that surface?
- Why did I move left?
- How can I tell how heavy that object is?
- Ditto ... how hard ...
- etc....

# **METHODOLOGICAL STANCE FOR SCIENCE**

---

In order to have a deep understanding of any ONE architecture, we need to understand

- the ‘surrounding’ space of possible architectures
- the states and processes they can and cannot support,
  - including the varieties of types of mental states and processes
- The trade-offs between different designs in different contexts.
- the variety of possible sets of **requirements** for such architectures (the niches)
- The different kinds of learning and development that are possible in different architectures
  - Gradual adaptation of skills
  - New chunks of capabilities
  - New ways of combining capabilities
- interactions between trajectories (evolutionary, individual, cultural) in ‘niche space’ and in ‘design space’, for a whole individual and for components.
- **What are the niches that drive evolution of flexibility of various kinds?**

Answering those questions will help us understand why humans, chimps, lions and crows are (largely) altricial, not precocial like deer, horses, chickens and insects.  
See this draft paper <http://www.cs.bham.ac.uk/research/cogaff/altricial-precocial.pdf>
- **Which architectures can support human-like capabilities?**

# Examples of architectural choices

---

## Choices for visual subsystem capabilities

- Variable diameter receptive fields
- Visual system maintains model across saccades and other movements
- Concurrent visual processes performing different tasks at various levels
  - Recording 3-D surface boundaries
  - Recording shape features (various kinds)
  - Recording and tracking moving objects (how many at a time?)
  - Searching for interesting novel patterns (things to learn)
  - Monitoring occurrences of 'salient' events: flashes, unexpected movements, appearance/disappearance of objects
- Visualising something that is not being perceived
  - What if:
    - The big object moves through the gap ?
    - Those two objects are moved closer ?
    - There is a box behind the brick ?

## Choices for auditory subsystem capabilities

- Support for detecting predators, prey, effects of actions, etc.
- Support for speech processing

## Choices regarding forms of representation:

numeric, qualitative, structural, histograms, maps, Trehub's 'retinoid'

# Varieties of deliberative mechanisms

---

These differ in various ways:

- the forms of representations (data-structures in virtual machines)
- the variety of forms available (e.g. logical, pictorial, rules, activation vectors)
- the algorithms/mechanisms available for manipulating representations
- the number of possibilities that can be represented simultaneously
- the depth of ‘look-ahead’ in planning
- the ability to represent future, past, or remote present objects or events
- the ability to represent possible actions of other agents
- the ability to represent mental states of others (linked to meta-management, below).
- the ability to represent abstract entities (numbers, rules, proofs)
- the ability to learn, in various ways

Some deliberative capabilities require the ability to learn new abstract associations, e.g. between situations and possible actions, between actions and possible effects.

Some require the ability to extend one’s ontology, e.g. to be able to think about acceleration, about electrical resistance, etc.

Some require the ability to acquire new syntactic forms used in composition of novel structures.

# Evolutionary pressures on perceptual and action mechanisms for deliberative agents

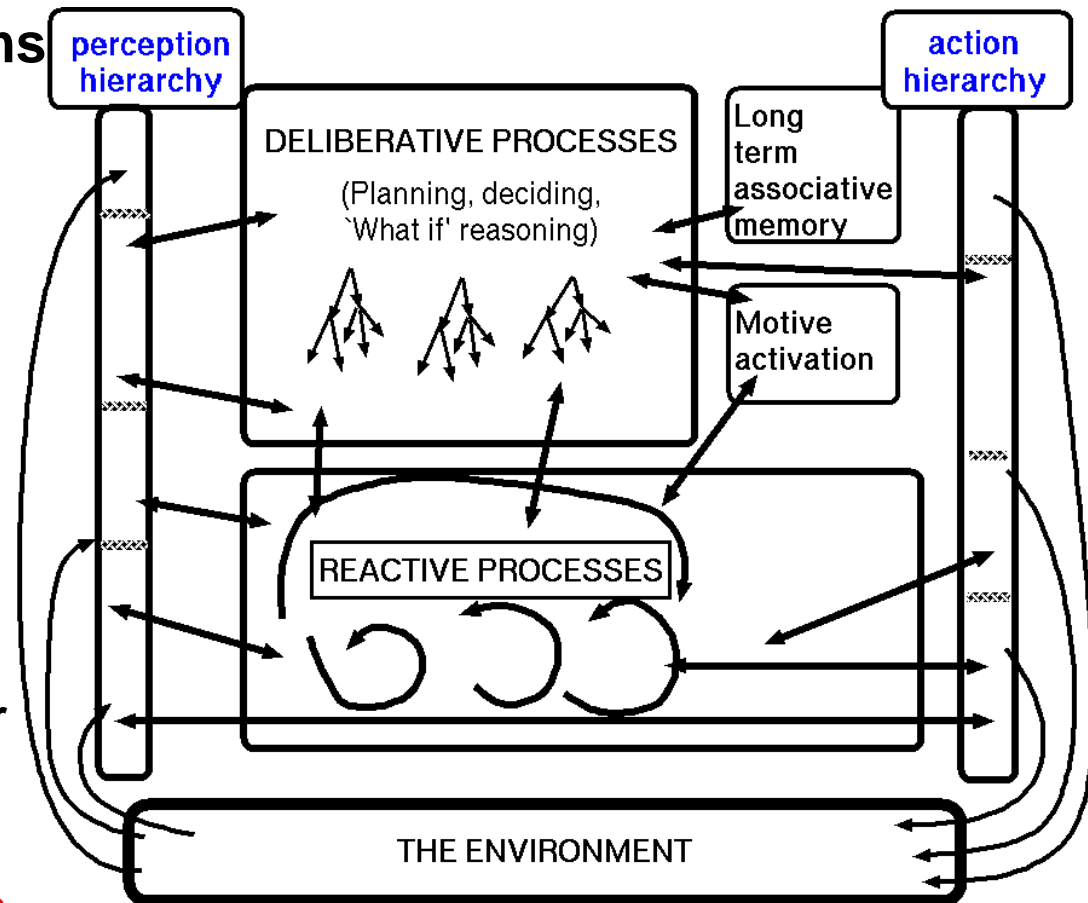
Layered central mechanisms co-evolved with

- new levels of perceptual abstraction (e.g. perceiving object types, abstract affordances),
- new mechanisms supporting high-level motor commands (e.g. “walk to tree”, “grasp berry”)

helping to meet requirements for deliberative processes.

Hence **taller, layered, perception and action towers** in the diagram.

We call that ‘multi-window’ perception and action, contrasted with Omega Architectures, which use only ‘peephole’ perception and action.



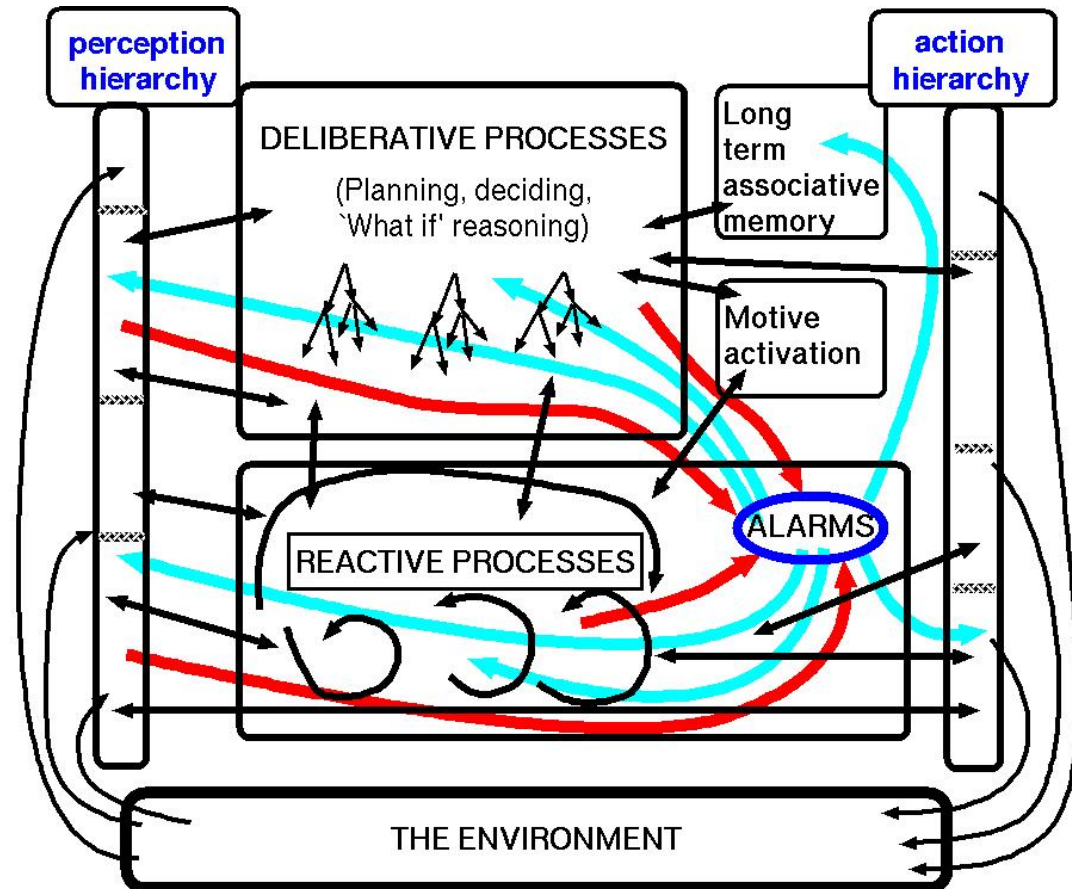
# A deliberative system may need an alarm mechanism

Inputs to an alarm mechanism may come from anywhere in the system, and outputs may go to anywhere in the system.

An alarm system can override, interrupt, abort, or modulate processing in other systems.

It can also make mistakes because it uses **fast** rather than **careful** decision making.

Learning can both extend the variety of situations in which alarms are triggered and improve the accuracy.



False positives and false negatives can result both from limitations in the learning mechanism and from features of the individual's history: as attested by many aspects of human emotion.

# Some alarms may need filtering

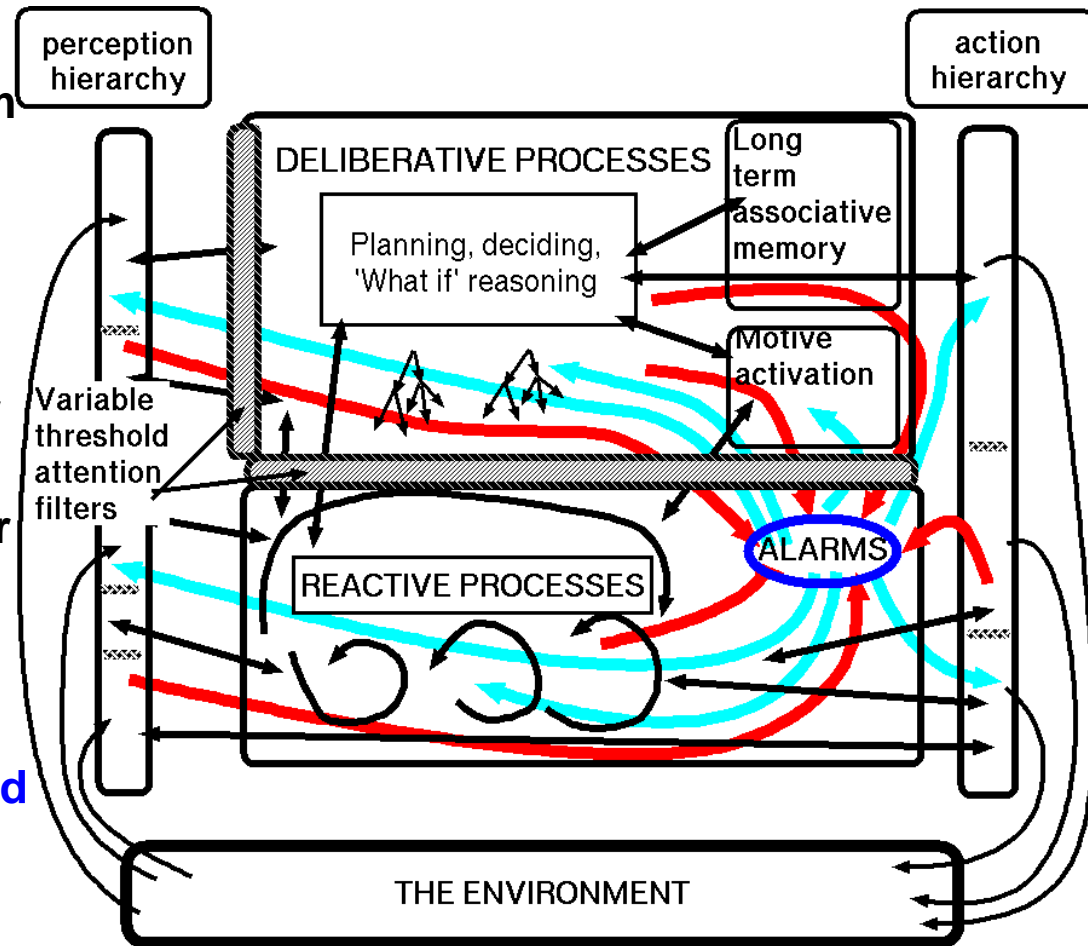
An alarm signal produced by an unintelligent reactive mechanism could disrupt some more urgent and important deliberative process.

In order to reduce that risk, attention filters with dynamically modulated thresholds, help suppress some alarms and other disturbances during urgent and important tasks.

Many human emotions are concerned with perturbances and limitations of attention filtering mechanisms, including some long term emotions, like grief

See

I.P. Wright, A. Sloman & L.P. Beaudoin, (1996), Towards a Design-Based Analysis of Emotional Episodes, *Philosophy Psychiatry and Psychology*.



# **Multi-window perception and action**

If multiple levels and types of perceptual processing go on in parallel, we can talk about

“multi-window perception”,

as opposed to

“peephole” perception.

Likewise, in an architecture there can be

multi-window action

or merely

peephole action.

In multi-window perception, perceptual processes operate concurrently at different levels of abstraction serving the needs of different cognitive processing layers.

Likewise multi-window action.

**CLAIM:**

The emphasis on recognition, localisation, moving and tracking, as opposed to **manipulation** of objects has distracted attention from understanding human-like vision and perception of spatial and causal structures (affordances).

**(Compare Freddy II the Edinburgh robot: 1973.)**

# Requirements for meta-semantic and meta-management competence

---

## Regarding others

- Thinking about their mental states – beliefs, preferences, goals, knowledge, memories
- Thinking about what they can see, achieve, etc.
- Thinking about how to influence
  - their actions
  - their thoughts decisions, etc

## Regarding self

- Thinking about one's mental states – beliefs, preferences, goals, knowledge, memories
- Thinking about what one can see, achieve, etc.
- Thinking about how to change one's sensory experiences
- Thinking about how to change one's reasoning, planning, action-control capabilities and strategies.

**Question:** what goals, preferences, values, etc. will PlayMate or Explorer have, and how will they change?

Will it **enjoy** some experiences or actions, and **dislike** others? Is this a requirement for learning?

# Requirements for adding linguistic competence

**Basic**

**Fluent**

**Speech**

**Text**

**Conjecture: linguistic competence builds on much pre-linguistic competence, but some of it is copied and transformed.**

# **MAIN Features of reactive organisms**

---

The main feature of reactive systems is that they **lack the core ability of deliberative systems**, namely

to represent and reason about phenomena that either do not exist or are not sensed

e.g.

**future possible actions,  
remote entities,  
the past, hidden items  
etc.**

- In principle a reactive system can produce any external behaviour that more sophisticated systems can produce (e.g. using huge collections of condition-action rules, where some of the conditions are internal)
- However, in practice there are constraints ruling this out, for instance the need for physical memories too large to fit on a planet.
- These constraints forced evolution to produce fully deliberative mechanisms in a subset of species
- Note: deliberative mechanisms have to be *implemented* in reactive mechanisms, in order to work: but that does not stop them having deliberative capabilities.

# PROTO-DELIBERATIVE SYSTEMS

---

Evolution also produced proto-deliberative species:

- In a reactive system (e.g. implemented as a neural net) some sensed states with mixtures of features can simultaneously activate two or more incompatible response-tendencies (e.g. fight and flee).
- In that case some sort of competitive mechanism can select one of the options, e.g. based on the relative strengths of the two sensory patterns, or possibly based on the current context (internal or external e.g. level of hunger or whether an escape route is perceived).

**Here alternative futures are represented and then a selection is made.**

**Some people call this deliberation.**

- However, such a system lacks most of the features of a **fully deliberative system** so we can call it a **proto-deliberative system**

**Going beyond reactive or proto-deliberative systems towards fully deliberative systems requires major changes in the architecture, though evolution may have got there by a collection of smaller, discrete, changes: we need to understand the intermediate steps.**

**Note: 'deliberative' and 'symbolic' are not synonyms. A purely reactive system may use symbolic condition-action rules (e.g. Nilsson's 'teleoreactive systems').**

# **Did Good Old Fashioned AI (GOF AI) fail?**

---

It is often claimed that symbolic AI and the work on deliberative systems failed in the 1970s and 1980s and therefore a new approach to AI was needed.

**THIS IS A COMPLETE MISDIAGNOSIS.**

**What actually happened was that symbolic AI research failed to fulfil *inappropriate* predictions made by researchers (some in symbolic AI) who had not understood the problems.**

This is equally true of all other approaches to AI: many of the problems are subtle, complex, and still not understood. E.g. how should perceived shape be represented?

See <http://www.cs.bham.ac.uk/research/cogaff/challenge.pdf>

The recent emphasis on **architectures** helps us think more clearly about combining **different sorts of components** with **different functional roles** (including reactive and deliberative subsystems) working together.

That is an essential step towards understanding (and perhaps eventually replicating) human capabilities.

**Another Anti-AI red herring: ‘Symbol-grounding theory’**

What we really need is ‘*symbol-attachment*’ theory.

See <http://www.cs.bham.ac.uk/research/cogaff/talks/#meanings>

# Representing what does not exist can be useful

**Deliberative mechanisms provide the ability to represent unsensed possibilities (e.g. possible actions, possible explanations for what is perceived).**

**One application of that is planning multi-step actions, including nested actions (unlike 'proto-deliberation', which considers alternative single-step actions, and can use simple neural net mechanisms).**

**Much, but not all, early symbolic AI (surveyed in Margaret Boden's 1978 book *Artificial Intelligence and Natural Man*) was concerned with deliberative systems (planners, problem-solvers, parsers, theorem-provers, concept-learners, analogy mechanisms,.....).**

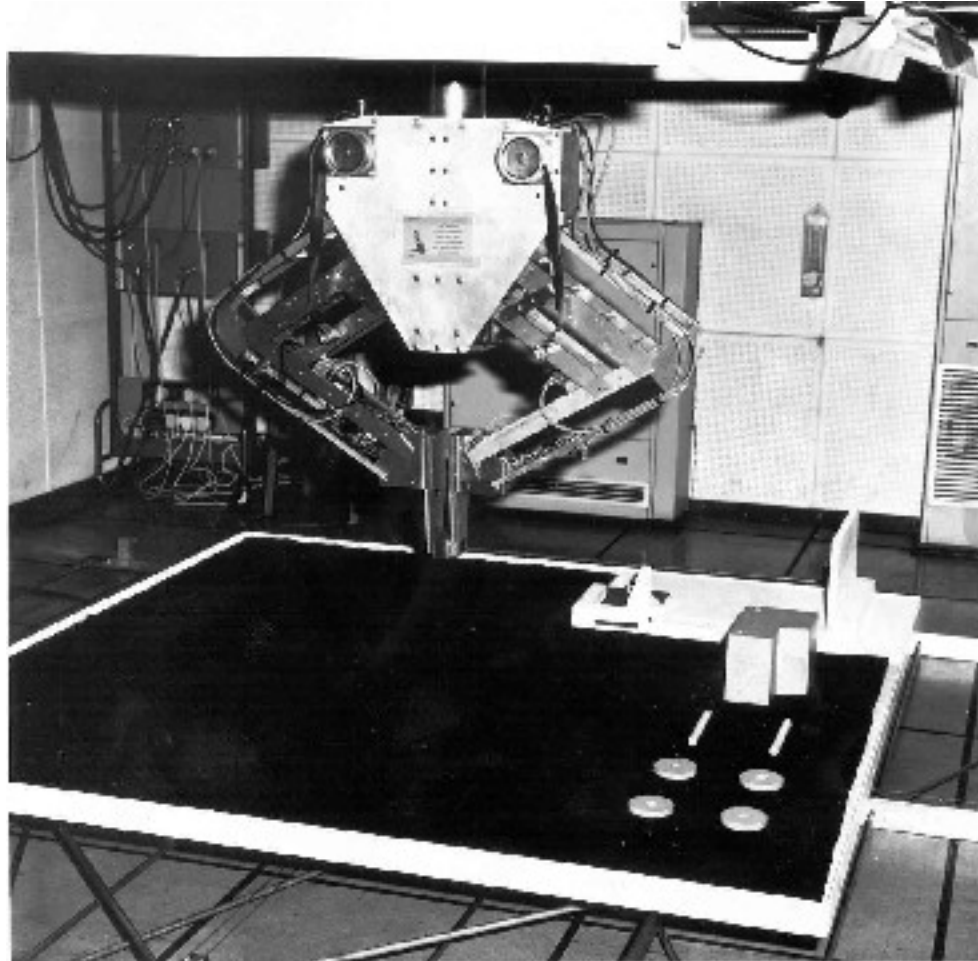
**There were also experiments with reactive systems: simple creatures that reacted to their needs, drives, and externally sensed phenomena, and possibly learnt in simple ways.**

**There are demo movies of a purely reactive symbolic simulated sheepdog herding sheep, and a hybrid deliberative/reactive one, with planning capabilities here:**

**<http://www.cs.bham.ac.uk/research/poplog/figs/simagent/>**

# FREDDY THE SCOTTISH ROBOT (1973)

---



**Freddy, developed in Edinburgh using mostly symbolic AI techniques, could assemble a few objects (like the toy car shown) from parts, which did not have to be arranged tidily as in the picture.**

# The pressure towards self-knowledge, self-evaluation and self-control

---

A deliberative system can easily get stuck in loops or repeat the same unsuccessful attempt to solve a sub-problem, or use thinking strategies with flaws.

- One way to reduce this is to have a parallel sub-system monitoring and evaluating the deliberative processes.

(Compare Minsky on “B brains” and “C brains” in *Society of Mind*, and Sussman’s HACKER.)

- We call this meta-management. It seems to be rare in biological organisms and probably evolved very late – to support altricial species.
- As with deliberative and reactive mechanisms, there are many forms of meta-management.

# Later, meta-management (reflection) evolved

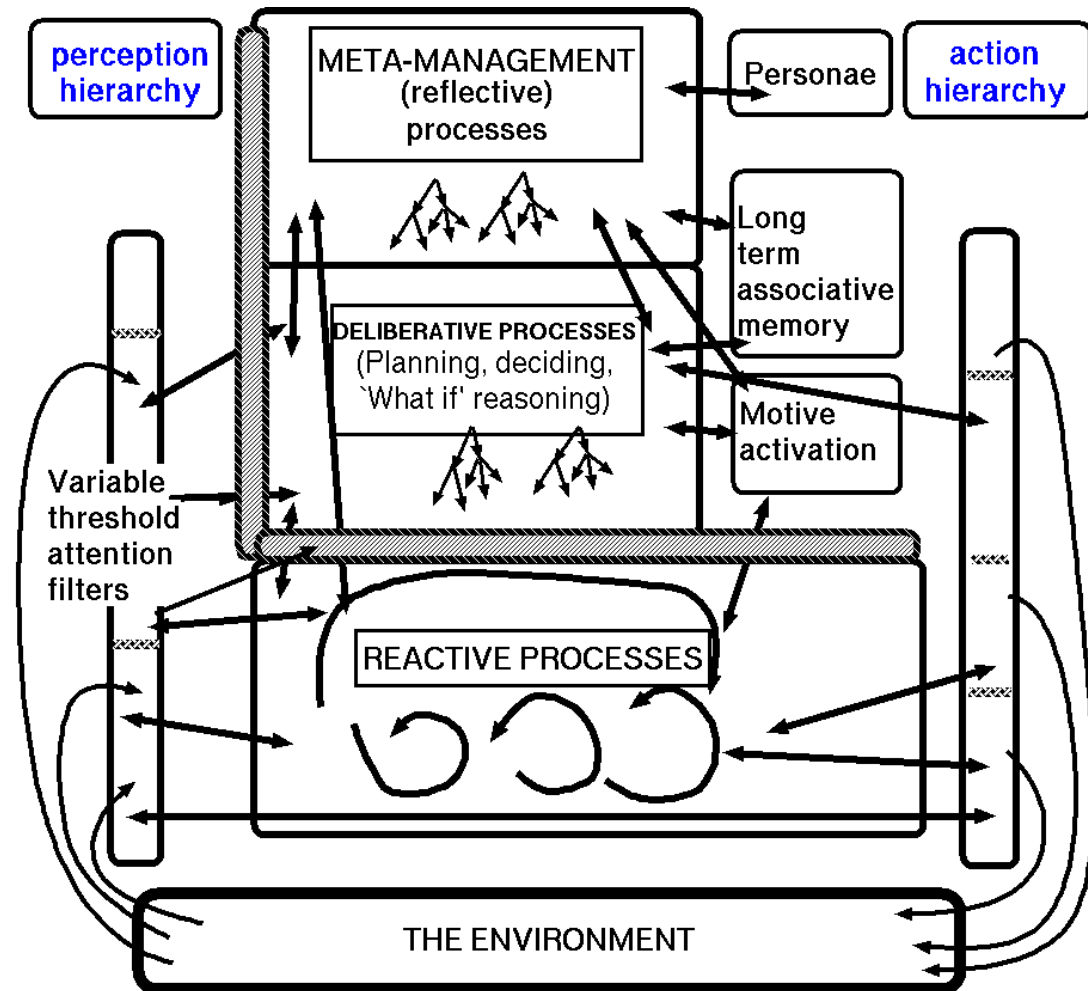
An addition to reactive and deliberative competence including both reactive and deliberative mechanisms.

Can be viewed as a generalisation of homeostasis.

Self monitoring, can include categorisation, evaluation, and (partial) control of internal processes.

Not just measurement.

The richest versions of this evolved very recently, and may be restricted to humans.



**Absence of meta-management can lead to stupid behaviour in AI systems, and in brain-damaged humans.**

See A.Damasio (1994) *Descartes' Error* (but watch out for the fallacies).

# **Inner and outer perception co-evolved**

---

## **Conjecture:**

**the representational capabilities that evolved for dealing with self-categorisation can also be used for other-categorisation, and vice-versa. Perceptual mechanisms may have evolved recently to use these those representational capabilities in percepts.**

**Example: seeing someone else as happy, or angry.**

# **Further steps to a human-like architecture**

---

## **CONJECTURE:**

**Central meta-management led to opportunities for evolution of**

– **additional layers in ‘multi-window perceptual systems’**

**and**

– **additional layers in ‘multi-window action systems’,**

**Examples: social perception (seeing someone as sad or happy or puzzled), and stylised social action, e.g. courtly bows, social modulation of speech production.**

**Additional requirements led to further complexity in the architecture, e.g.**

– **‘interrupt filters’ for resource-limited attention mechanisms,**

– **more or less global ‘alarm mechanisms’ for dealing with important and urgent problems and opportunities,**

– **socially influenced store of personalities/personae**

**All shown in the next slide, with extended layers of perception and action.**

# More layers of abstraction in perception and action, and global alarm mechanisms

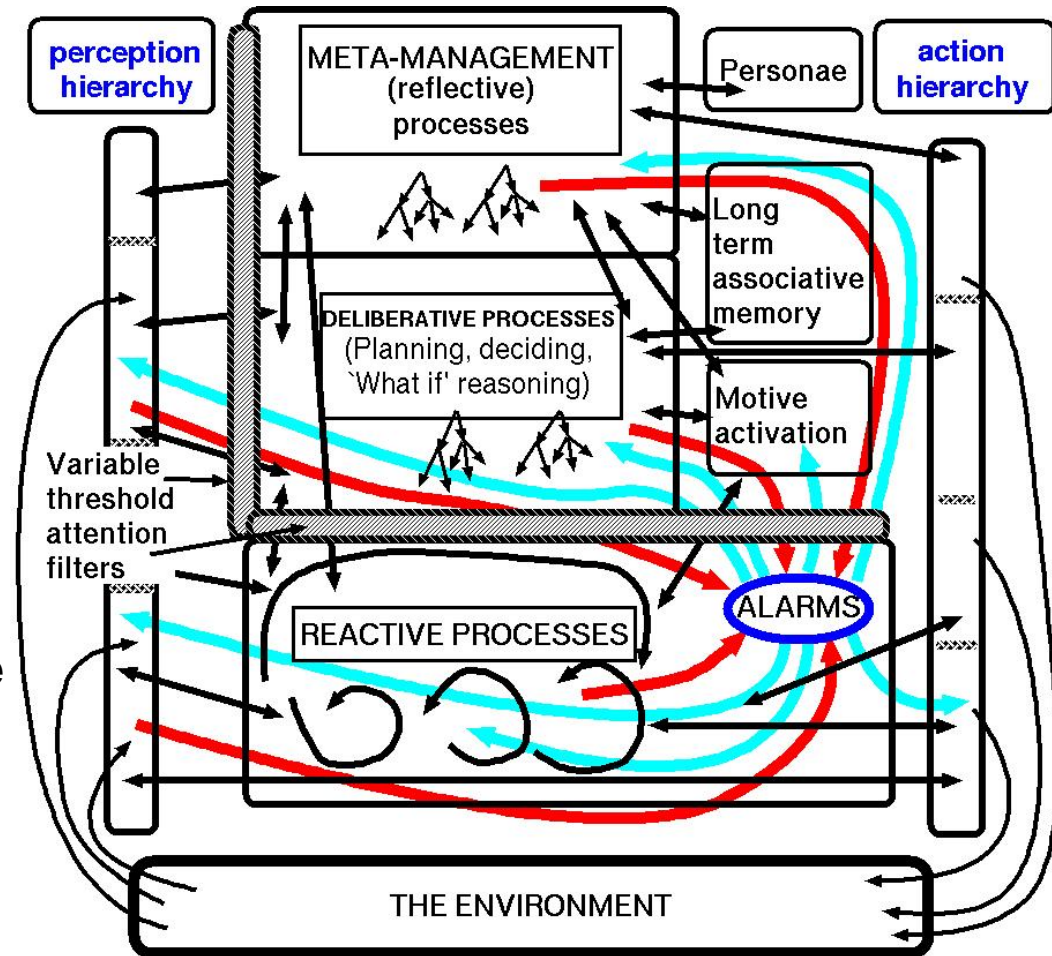
This conjectured architecture (H-Cogaff) could be included in robots (in the distant future).

Arrows represent information flow (including control signals)

If meta-management processes have access to intermediate perceptual databases, then this can produce self-monitoring of sensory contents, leading robot philosophers with this architecture to discover “the problem(s) of Qualia?”

‘Alarm’ mechanisms can achieve rapid global re-organisation.

Meta-management systems need to use **meta-semantic** ontologies: they need **the ability to refer to things that refer to things**.



## Implications continued ....

---

- **Many varieties of learning and development**  
(E.g. “skill compilation” when repeated actions at deliberative levels train reactive systems to produce fast fluent actions, and action sequences. Needs spare capacity in reactive mechanisms, (e.g. the cerebellum?). We can also analyse development of the architecture in infancy, including development of personality as the architecture grows.)
- **Conjecture: mathematical development depends on development of meta-management – the ability to attend to and reflect on thought processes and their structure, e.g. noticing features of your own counting operations, or features of your visual processes.**
- **Further work may help us understand some of the evolutionary trade-offs in developing these systems.**  
(Deliberative and meta-management mechanisms can be very expensive, and require a food pyramid to support them.)
- **We can see how philosophical thoughts (and confusions) about consciousness are inevitable in intelligent systems with partial self-knowledge.**

For more see papers here: <http://www.cs.bham.ac.uk/research/cogaff/>

# What is an architecture?

---

AI used to be mainly about **algorithms** and **representations**.  
Increasingly, during the 1990s and onward it has been concerned with the study of **architectures**.

An architecture includes:

- **forms of representation,**
- **algorithms,**
- **concurrently processing sub-systems,**
- **connections between them.**

Note: Some of the sub-systems may themselves have complex architectures.

Note: Don't confuse **components** and **capabilities**  
(beware of 'emotion' boxes.)

**Architectures need not be fixed:**

they can develop  
especially in altricial species.

Human information processing architectures continue developing as new sub-ontologies are learnt (e.g. chemistry, physics, biology, computing), as new languages are learnt (natural and formal), and as new types of skills are learnt (e.g. athletic skills, musical skills, artistic skills.)

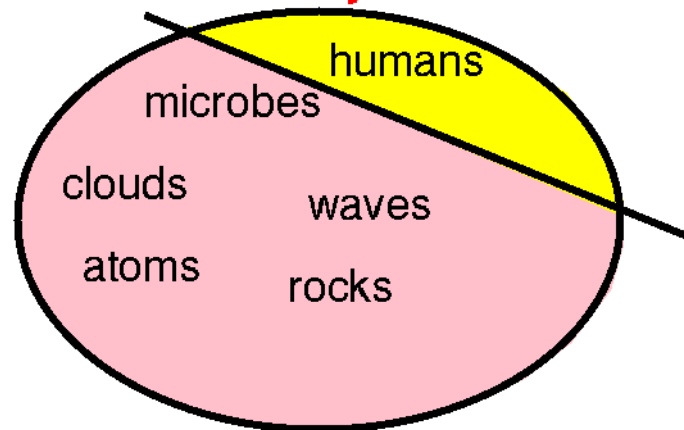
# There's No Unique Correct Architecture

Some tempting **wrong** ways to think about consciousness:

1. There's no **continuum** from non-conscious to fully conscious beings



2. It's not a **dichotomy** either



**Both 'smooth variation' and a single discontinuity are poor models. We need to understand tradeoffs and discontinuities.**

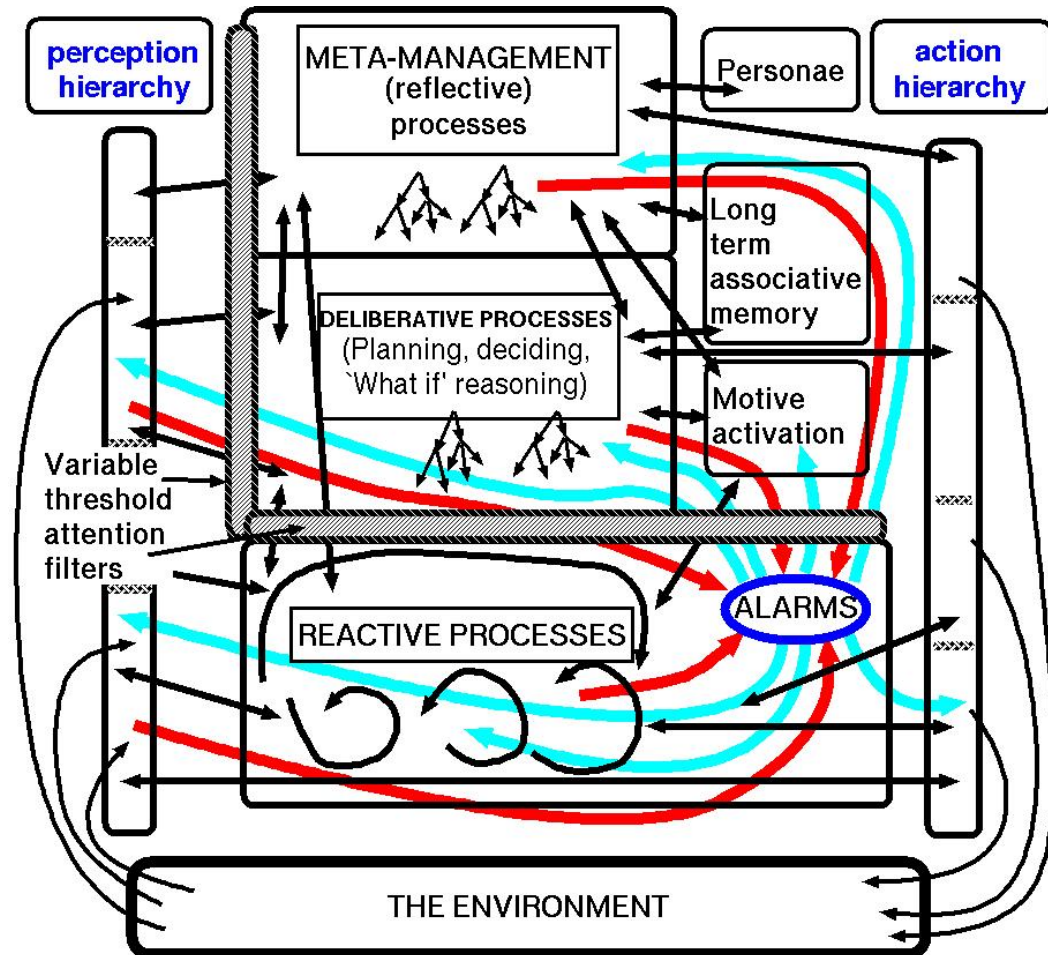
# Putting Pieces Together: H-CogAff

A postulated architecture for human-like systems.

**MANY** kinds of things going on in parallel, doing different things, concurrently – some discrete, some continuous, some low-level, some high level, some concrete, some abstract, lots of interactions, ..... (a very long term project)

Explained in more detail elsewhere.

We shall not be able achieve all of this, but we can analyse the requirements in PlayMate and Explorer scenarios for some of the components.



# A corollary of all the parallelism in H-CogAff

We must kill this silly, but often recommended model

**SENSE ⇒ DECIDE ⇒ ACT**

which ignores architectures with multiple concurrent components.