TALK FOR SHEFFIELD SEMINAR 20-04-2005, and READING SEMINAR 19-05-2005

Aaron Sloman

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

School of Computer Science

The University of Birmingham

Putting the pieces of a mind together.

Varieties of functionality found in nature

e.g. the precocial-altricial spectrum

Crows that understand hooks

children that do not

The space of architectures

Varieties of components

How architectures develop

We need an ontology for architectures (designs) and requirements (niches)

Misunderstanding requirements: towards really seeing.

Online at http://www.cs.bham.ac.uk/research/cogaff/talks/reading.pdf

Related slides available at http://www.cs.bham.ac.uk/research/cogaff/talks/

THANKS

I am very grateful to
the developers of Linux
and other free, open-source,
platform-independent, software systems.

LaTex was used to produce these slides.

Diagrams are created using tgif, freely available from

http://bourbon.cs.umd.edu:8001/tgif/

Demos are built on Poplog

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

APOLOGIES

- This talk is not about some specific piece of research and its results.
- It is about a long term research programme —

  Grand Challenge 5 Architecture of Brain and Mind in the UKCRC list of research grand challenges.
  See http://www.ukcrc.org.uk/
- All I can do in a short summary is present a selection of the issues.
- More details are available here

  http://www.cs.bham.ac.uk/research/cogaff/talks/
  http://www.cs.bham.ac.uk/research/cogaff/gc/

  Or email me: A.Sloman@cs.bham.ac.uk
- I'll give a quick summary of issues and discuss a selection in detail, with illustrative demos and videos, depending on audience.

NOTE: Slides marked 'skip' will be skipped during the presentation, but may be useful for people reading the slides.

Summary of full talk: part 1

- Current ways of studying (animal, human and robot) minds are
  - too fragmented
  - too riddled by turf wars
  - too much influenced by prejudice (what people would like to be true)
  - based on inadequate notions of science and explanation
  - based on too little data in forms that are too restricted

- Examples:
  - bad theories about emotions, about vision, about meaning (e.g. symbol-grounding),
  - confused concepts treated as well understood, e.g. emotion, consciousness, learning...
  - theories/models/explanations that don't 'scale out' (fit into a larger context)
    ('Scaling out' is more important than 'scaling up' — for a science of mind.)
  - Distorting history: e.g. claiming that 'symbolic AI failed' (it has barely begun).
  - Fads and fashions instead of theories e.g. 'everything is reactive'...

- We can remedy this by working out the implications of these facts:
  - minds DO things: they are constantly active machines
  - there is not just one kind of mind: very many exist in nature, even among humans: young, old, normal, damaged, ancient, modern (industrialised)
  - all organisms are information processors
  - evolution is far ahead of our understanding
  - all complex designs involve complex trade-offs
  - new evolutionary designs do not simply throw away old solutions, but build on them:
    humans share much with much older species

Summary of full talk: part 2

- **This project requires contributions from many disciplines – and they will be changed by contributing to the project.**

- **It can also build on common sense**

  A new science of mind need not throw away the rough-hewn concepts of ordinary language, and the vast amount of folk-knowledge we use every day (mostly unconsciously, much embedded in our use of language), but can use them as stepping stones to a richer, deeper, collection of ways of thinking about what sorts of machines we are, and might be.

    Compare the way physics, as deep explanatory science, builds on and explains 'folk physics' instead of throwing it all away.

- **A major contribution from Computer Science, AI and Software engineering is new understanding of levels of abstraction and their relationships:**

  Layers of virtual machines running on physical machines.
  (But our understanding is still rudimentary).

- **Understanding ourselves (and other animals), i.e. understanding how we work, is partly an exercise in designing working systems – not just a matter of collecting, correlating, summarising, organising, observations**

The context:
Structures and structural change pervade biology

A few examples:

- **Cellular division and repair**
- **Growth of individual organisms**
- **Ingestion and digestion of food and use of the materials**
- **Nest building**
- **Reproduction**
- **Social processes**
- **Evolutionary change .... and many more**

Some of the structures and processes are physical.
Others involve virtual machines

E.g. concept formation is not a physical process, even if implemented in physics.

Features of biological virtual and physical machines:

A high degree of parallelism.

Many kinds of processing of information, of many kinds:

sensing **acting** **....**	**transforming** **adapting** **...**	**controlling** **communicating,** **....**

SOME KEY IDEAS AND QUESTIONS

In addition to physical growth –
biological organisms also grow
information-processing architectures
which are virtual machines,
not physical machines

What does this mean?

How does it happen?

How much of it is genetically determined?

How much is controlled by the environment?

How much is controlled by the individual?

How much is controlled by the culture?

PROBLEM

Unfortunately

People pick up one or two relatively simple ideas and run with them.

E.g. logic, artificial neural nets, GAs, reactive behaviours, situatedness, 'swarm intelligence', dynamical systems

And some say:
"Let's give up design,
or attempts to understand,
and hope for emergence
of things we can use."

DO YOU RECOGNISE YOURSELF?

SKIP Some old ways to study minds

There are many ways to study emotions and other aspects of human minds:

- **Reading plays, novels, poems** will teach you much about how people who see, act, have emotions, moods, attitudes, desires, etc. think and behave, and how others react to them — because many writers are very shrewd observers!
- **Studying ethology** will teach you about how mental phenomena, including cognitive capabilities vary among different animals.
- **Studying psychology** will add much extra detail concerning what can be triggered or measured in laboratories, and what correlates with what.
- **Studying developmental psychology** can teach you how the states and processes in infants differ from those in older children and adults.
- **Neuroscience** teaches us about physiological brain mechanisms that support and modulate mental states and processes, and are modulated by them.
- **Studying therapy and counselling** can teach you about ways in which things can go wrong and do harm, and some ways of helping people.
- **Studying philosophy** (with a good teacher) may help you discern muddle and confusion in attempts to say what minds are and how mental states and processes differ from one another and from physical states and processes.

Another, less familiar, way complements and enriches those ways.

SKIP A newer way: do some engineering design

Suppose you had to design animals (including humans) or robots capable of living in various kinds of environments, including environments containing other intelligent systems.

What sorts of information-processing mechanisms, including control mechanisms, would you need to include in the design, and how could you fit all the various mechanisms together to produce all the required functionality, including:

- **perceiving (using various sensory systems),**
- **learning (concepts, forms of representation, facts, generalisations, skills, ...)**
- **acquiring new motives, values, standards, preferences, ideals ...)**
- **enjoying some activities and states and disliking others,**
- **selecting between conflicting motives,**
- **planning, executing plans, planning how to plan, ...**
- **reacting to dangers and opportunities,**
- **communicating in various ways (including collaborating, competing and fighting)**
- **reproducing, and so on...**

If we combine this "design standpoint" with the previously listed ways to study mental phenomena, we can learn much about all sorts of mental processes: what they are, how they can vary, what they do, what produces them, whether they are essential or merely by-products of other things, how they can go wrong, etc.

The result could be both deep new insights about what we (and other animals) are, and important practical applications.

The design-based approach – too fragmented now

The design-based approach is not new: over the last half century, researchers in Computational Cognitive Science, and in Artificial Intelligence have been pursuing it.

- **Because the work was so difficult and because of the pressures of competition for funding and other aspects of academic life (e.g. lack of time for study), the field fragmented, and as more people became involved the research community became more fragmented, with each group investigating only a small subset of the larger whole, and talking only to members of that group.**
- **Deep, narrowly focused, research on very specific problems is a requirement for progress, but if everybody does only that, the results will be bad.**
  - **People working on natural language without relating it to studies of perception, thinking, reasoning, and acting may miss out on important aspects of how natural languages work.**
  - **Likewise those who study only a small sub-problem in perception may miss out ways in which the mechanisms they study need to be modified to fit into a larger system.**
  - **The study of emotions also needs to be related to the total system.**

The European Community's recent initiative in 'Cognitive Systems' is an attempt to remedy this by requiring researchers to think about integrated multi-component systems.
One of the projects to be funded (including Birmingham) under that initiative is described here:
http://www.cs.bham.ac.uk/research/projects/cosy/

A UK grand challenge proposal to put all the pieces together again in a long term research programme is described here http://www.cs.bham.ac.uk/research/cogaff/gc/

The need for integration

Over the last half-century there has been much fragmentation, within each of: AI, psychology, neuroscience — most researchers focus only on a limited sub-field, e.g.

- **vision (usually low-level vision nowadays)**
- **language (text, speech, sign-language)**
- **learning (many different kinds)**
- **problem solving**
- **planning**
- **mathematical reasoning**
- **motor control**
- **emotions**

  etc....

Can the systems produced in each sub-field work fruitfully with systems produced in other sub-fields, within a common architecture?

Scaling up vs Scaling out

If our aim is to understand and model natural systems, including humans, the need to 'scale out' is more important than the need to 'scale up'

There is no guarantee that a technique, or form of representation, or algorithm, etc. that works for an isolated task will also work when that task has to be integrated with many other kinds of functionality in an integrated system.

AI techniques that 'scale up' very well within a particular application domain, e.g. path planning, may not 'scale out' to support anytime planning or reasoning about planning, or cooperative planning using natural language, or planning in a visually perceived constantly changing context.

Human abilities generally do not scale up
(Donald Michie labelled this 'the human window').
We are defeated by combinatorics and by structural complexity.
But human abilities 'scale out' and interact fruitfully:
e.g. what you see can help you understand
words you hear and vice versa. (McGurk effect)
And your visual spatial competence can help you think about
abstract mathematics, e.g. transfinite ordinals, category theory.

Putting the pieces back together:

- We need to understand and model brains/minds as integrated working systems functioning at different levels of abstraction, including
  - Physiological properties of brain mechanisms (how many different sub-types are there?)
  - Neural information processing functions
  - 'Higher level' cognitive and affective functions of many sorts, implemented in older mechanisms.
  - Behaviours of complete agents (including social behaviours).
- This requires us to understand how the different levels, and the different components at each level, combine to form an integrated functioning system
  - some levels implementing others,
  - some sub-systems cooperating with or competing with others
- We need to understand principles of operation at different levels rather than always merely trying to mimic low level biologicall details.

  Compare: the understanding of software engineers and digital electronic engineers, or physicists.
- At each level different kinds of functionality are integrated.
- We still lack good characterisations of requirements for architectures, mechanisms, formalisms, at all levels, a prerequisite for producing good designs

What sort of architecture? Tentative example: H-Cogaff

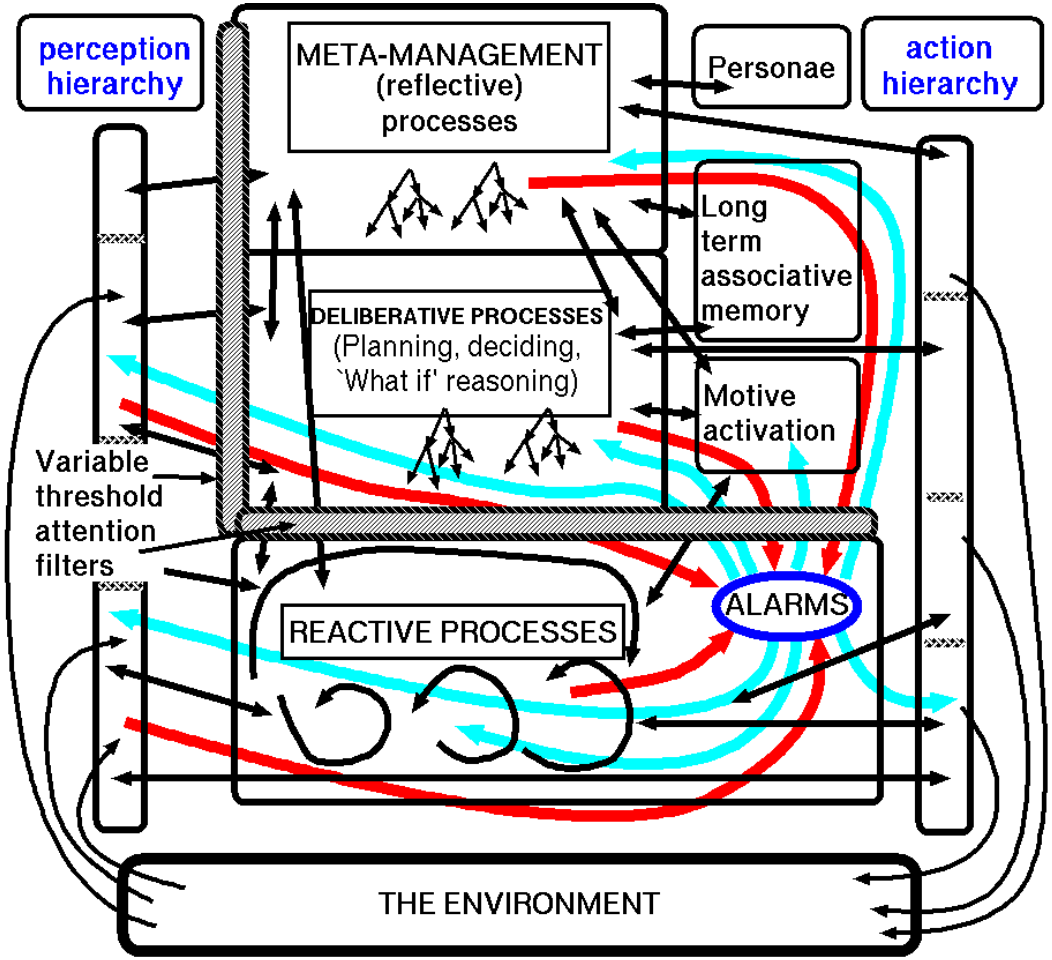
This diagram, representing schematically one possible type of architecture, is explained in more detail later.

The main point is that there are different subsystems operating concurrently, performing different sorts of tasks.

Some (near the bottom of the diagram) are evolutionarily old, and similar to many other kinds of animals.

The other subsystems are newer and do tasks that far fewer animals can perform.

Both perception and action operate concurrently at different levels of abstraction, in relation to different central sub-systems.



Contrast 'peephole perception' and 'peephole action', described later.

Using factual material

- **One problem is identifying what needs explaining.**

  Too often people observe only what their theories deem relevant, or collect only information that their statistical tools can process.

- **A scenario-based approach can help to overcome that limitation**

  by collecting and analysing very many real scenarios, organised according to their similarities and differences and ordered by complexity
  e.g. (of mechanisms, of information, of architectures, of representations needed).

- **Examples: collect and study videos of animals and children:**
  - **Betty, the new caledonian crow, surprised researchers at the Oxford University Zoology department when she displayed an ability to make a hook out of a straight piece of wire, in order to fish a bucket containing food out of a tube: (http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm)**
  - **An 18 month old child attempts to join two parts of a toy train by bringing two rings together instead of a ring and a hook, and showing frustration and puzzlement at his failure. (http://www.cs.bham.ac.uk/~axs/fig/josh34_0096.mpg)**
    A few weeks later he was able to solve the problem: what had changed?
  - **If time: video of the child playing with trains on the floor about a year later.**

- **Supplement observed scenarios with a large collection of analytical scenarios: compare Piaget**
  See also http://www.cs.bham.ac.uk/research/cogaff/targets.html
  http://www.cs.bham.ac.uk/~axs/polyflaps

Science or Engineering?

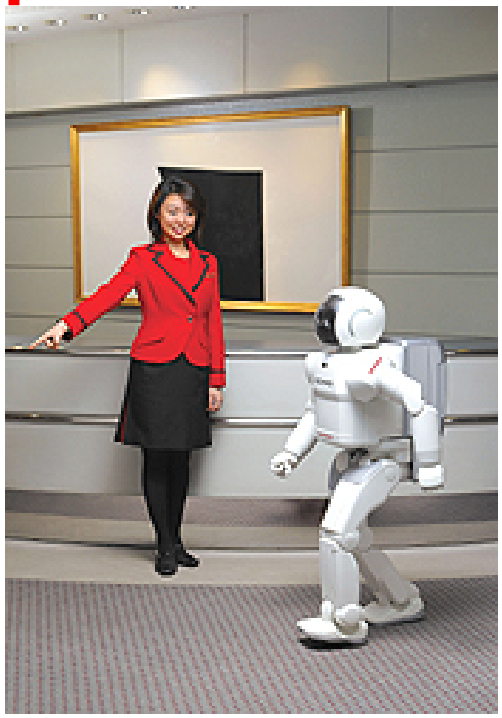
This is primarily a scientific challenge, not an applications challenge aimed at producing some useful new machines.

But the research has two aspects, theoretical and practical, which inform each other.

POTENTIALLY THERE ARE MANY APPLICATIONS – BUT THEY ARE NOT THE MAIN MOTIVATION.

The engineering goal of getting a machine to play chess as well as the best human players has been achieved, but not the scientific goal of clarifying requirements and designs for a machine that understands what it is doing when it plays chess, and can describe its strategy, explain things to a weaker player, etc.

Impressive robots made by Honda and Sony



THE STATE OF THE ART IN 2002

http://world.honda.com/news/2002/c021205.html

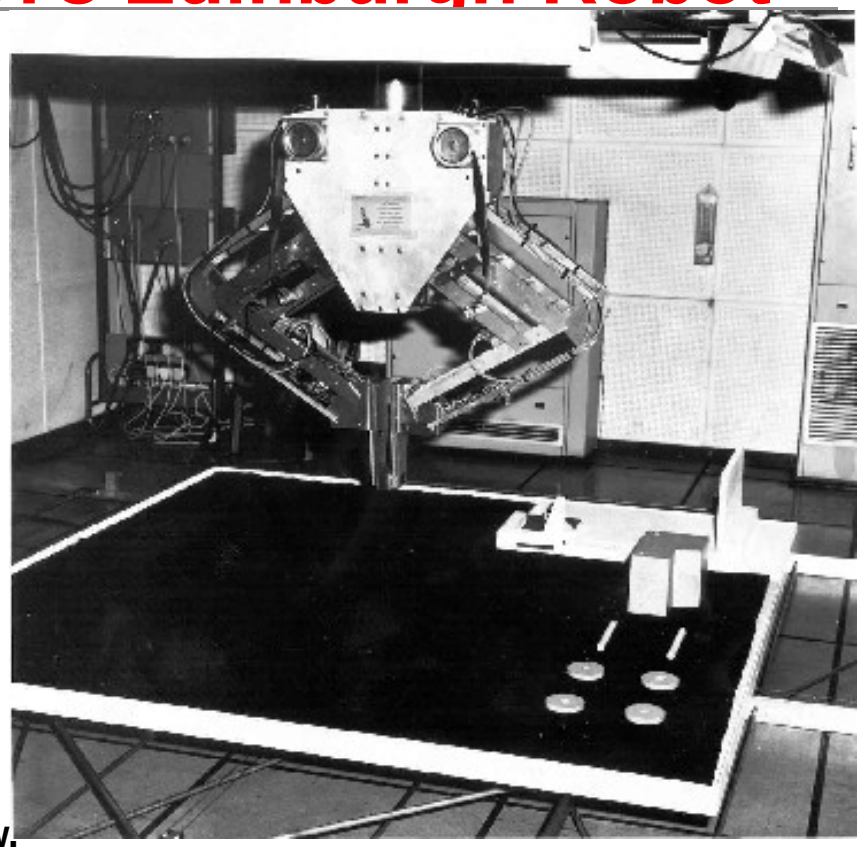
http://www.aibo.com/

In both cases the engineering is very impressive. But present day robots look incompetent if given a task that is even slightly different from what they have been programmed to do – unlike a child or crow or squirrel.

Mostly they have purely reactive behaviours, lacking the deliberative ability to wonder 'what would happen if...'.

They also have very little or no self-knowledge or self-understanding, e.g. about their limitations, or why they do things as they do.

Compare Freddy the 1973 Edinburgh Robot



Some people might say that apart from wondrous advances in mechanical and electronic engineering there has been little increase in sophistication since the time of Freddy, the 'Scottish' Robot, built in Edinburgh around 1972-3.

Freddy II could assemble a toy car from the components (body, two axles, two wheels) shown. They did not need to be laid out neatly as in the picture.

However, Freddy had many limitations arising out of the technology of the time.

E.g. Freddy could not simultaneously see and act: partly because visual processing was extremely slow.

Imagine using a computer with 128Kbytes RAM for a robot now.

There is more information on Freddy here

http://www.ipab.informatics.ed.ac.uk/IAS.html

http://www-robotics.cs.umass.edu/ pop/VAP.html

In order to understand the limitations of robots built so far, we need to understand much better exactly what animals do: we have to look at animals (including humans) with the eyes of (software) engineers.

What an organism or machine can do with information depends on its architecture

Not just its physical architecture – its information processing architecture.

This may be a virtual machine, like

- **a chess virtual machine**
- **a word processor**
- **a spreadsheet**
- **an operating system (linux, solaris, OS X, windows)**
- **a compiler**
- **most of the internet**

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

E.g. beware of hypothesised 'emotion' boxes, where a possible state is confused with a mechanism.)

An architecture can develop over time

especially in altricial species

(though parallel operation of new components may be limited)

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

METHODOLOGICAL PREREQUISITE

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

- the 'surrounding' space of information processing architectures
- the states and processes they can 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)
- interactions between trajectories (evolutionary, individual, cultural) in 'niche space' and in 'design space'.

- Which architectures can support human-like capabilities?

  Our ideas about this still have many gaps

- What are the niches that drive their evolution and require their variability?

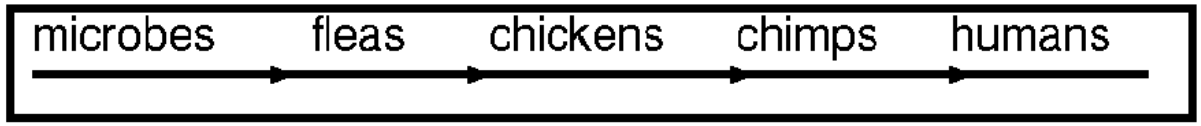
  Answering those questions will help us understand why humans, chimps, lions and crows are (largely) altricial, not precocial like deer, horses, chickens and insects.

  See this draft paper http://www.cs.bham.ac.uk/research/cogaff/altricial-precocial.pdf

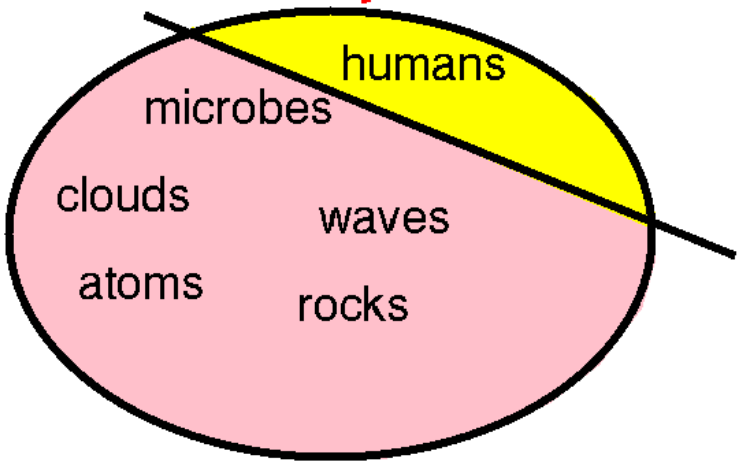
There's No Unique Correct Architecture

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.

Architectural challenges

One requirement for progress is specification of a virtual machine architecture that can combine many known kinds of human capabilities, including

- **evolutionarily very old reactive mechanisms**
- **newer deliberative mechanisms and**
- **biologically rare reflective, meta-management mechanisms with meta-semantic capabilities (the ability to represent processes in things that themselves represent other things, unlike rocks, trees, levers, wheels, blocks, ...).**

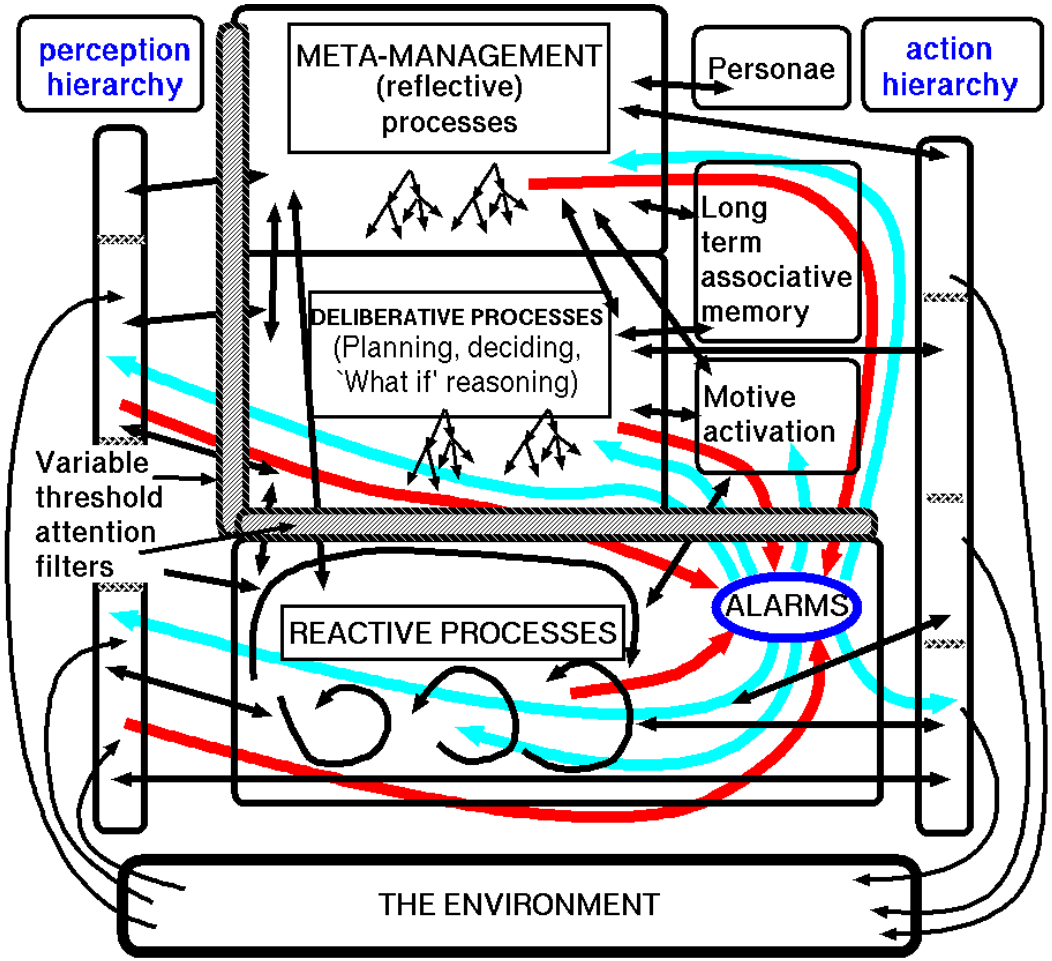
Papers and presentations in the Cognition and Affect project provide more detailed analyses of these architectural features, illustrated on the next slide. See

http://www.cs.bham.ac.uk/research/cogaff/
http://www.cs.bham.ac.uk/research/cogaff/talks/

Another glimpse of H-CogAff

A postulated architecture for human-like systems, explained in more detail later.

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)



We must kill the silly, but often recommended model:

SENSE ⇒ DECIDE ⇒ ACT

which ignores architectures with multiple concurrent components.

We need a generative ontology for architectures

'CogAff' is our label, not for an architecture (like 'H-Cogaff'), but for a way of specifying architectures – in terms of which sorts of components they include and how they are connected:
H-Cogaff is a special case of the schema.

Think of a grid of co-evolved types of sub-organisms, each contributing to the niches of the others, each performing different functions, using different mechanisms, etc.

We could add lots of arrows between boxes indicating possible routes for flow of information (including control signals) – in principle, mechanisms in any two boxes can be connected in either direction.

However, not all organisms will have all the kinds of components, or all possible connections.

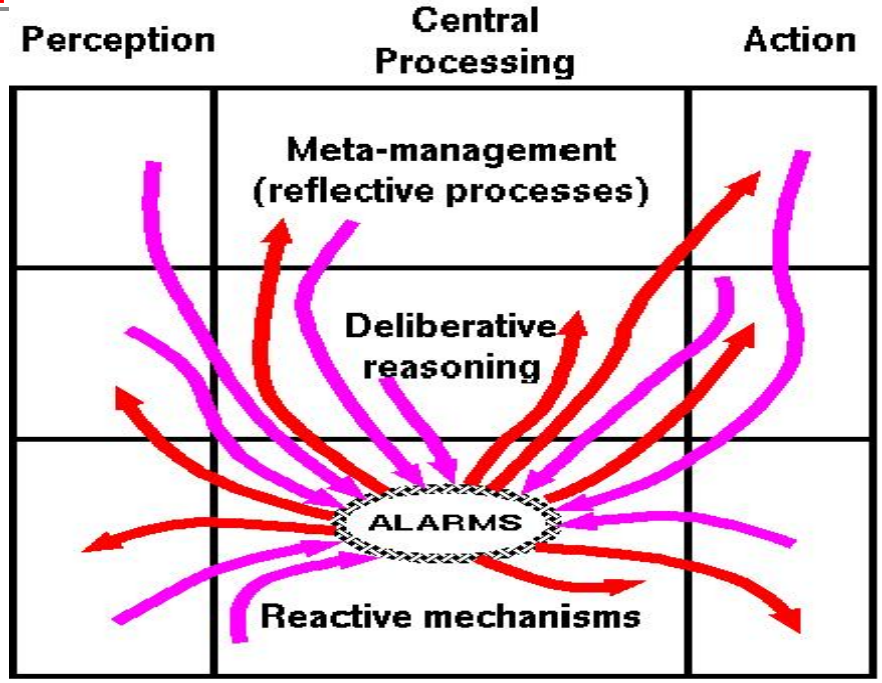
Perception	Central Processing	Action
	Meta-management (reflective processes) (newest)	
	Deliberative reasoning ("what if" mechanisms) (older)	
	Reactive mechanisms (oldest)	

E.g. insects are purely reactive, and perhaps also all reptiles and fish. A few species have deliberative capabilities in a simple form and perhaps even fewer have meta-management. Many kinds need "alarm" mechanisms.

As processing grows more sophisticated, so it can become slower, to the point of danger

REMEDY: FAST, POWERFUL, "GLOBAL ALARM SYSTEMS"

Resource-limited alarm mechanisms must use fast pattern-recognition and will therefore inevitably be stupid, and capable of error!



Many variants are possible. E.g. purely innate, or trainable.

E.g. one alarm system or several?

(Brain stem, limbic system, ...???)

See Cogaff papers and talks

http://www.cs.bham.ac.uk/research/cogaff/

http://www.cs.bham.ac.uk/research/cogaff/talks/

Many different kinds of emotional states can be based on such an alarm system, depending on what else is in the architecture.

Don't confuse the alarms (and emotions they produce) with the evaluations that trigger them, or the motives, preferences, policies, values, attitudes that have different sorts of functional roles – different sorts of control functions (including conditional control in many cases).

The ‘five Fs’

What we have called ‘alarm mechanisms’ may trigger behaviours often referred to as ‘the four Fs’, though there are at least five:

- **fleeing**
- **fighting**
- **feeding**
- **freezing**
- **reproducing**

(The usual list does not include ‘freezing’ – often the best response to danger or uncertainty.)

In humans they can trigger far more complex and subtle processes including deliberative and meta-management processes (e.g. reasoning anxiously about whether it would be wise to continue pursuing one’s current goal).

Compare: A simple (insect-like) architecture

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

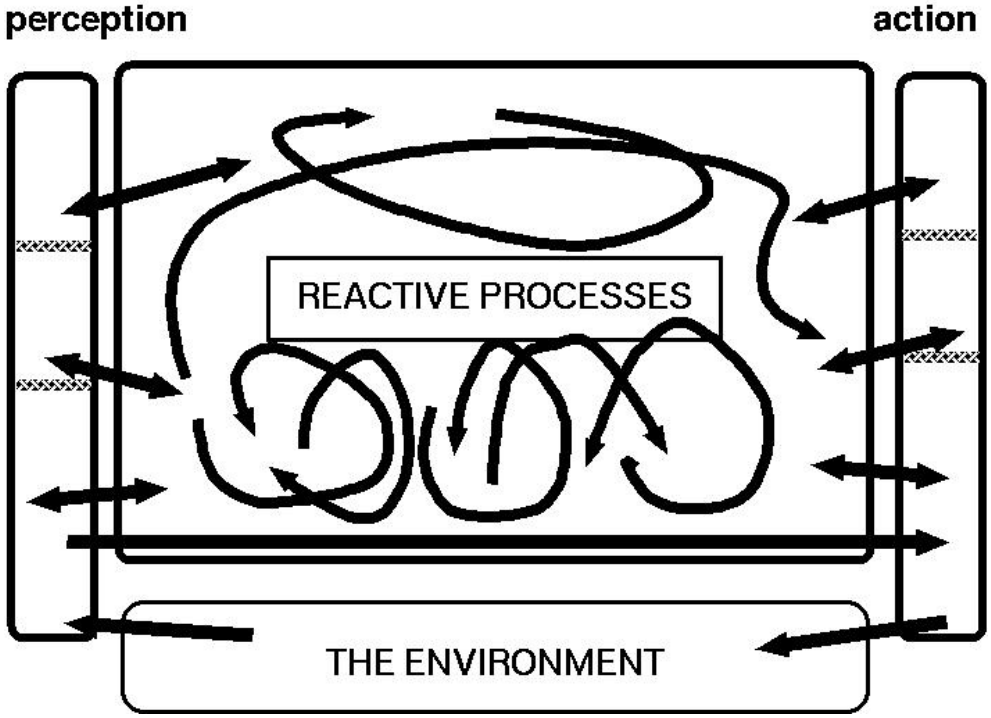
(But see proto-deliberation, later.)

It simply reacts: internally or externally.

Several reactive sub-mechanisms may operate in parallel.

Processing may use a mixture of analog and discrete mechanisms.

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.



Purely reactive biological species are precocial: they have large amounts of genetically determined capabilities, though minor environmentally driven adaptations are possible.

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 (GOFAI) 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 happend was that symbolic AI reseach 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

For many years AI research focused mainly on algorithms and representations.

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.

How does meaning get into the architecture?

Any organism, robot, or control system needs to acquire and use information about its environment and usually also about itself.

How can internal structures (symbols, neurons, networks of symbols or neurons) or internal processes, whether symbolic or not, be about anything (have intentionality, reference, sense, meaning, denotation, connotation,.....)?

This old philosophical problem — to which there are empiricist (e.g. Locke, Berekely, Hume,) and non-empiricist (e.g. Kant) answers — was rediscovered by AI critics and researchers, who reinvented concept empiricism and called it 'symbol-grounding' theory, sometimes used as an anti-AI weapon, when in fact it's a red-herring!

Extreme 'Symbol-grounding' theory: concepts are derived bottom-up by abstracting from experience of instances.
Kant (1781): you can't have experiences without having concepts to start with.

20th century philosophers of science (e.g. Carnap, Tarski) showed how meanings of theoretical terms in science (e.g. 'electron', 'quark', 'gene') come mainly from structural properties of theories using them (compare Tarskian semantics) augmented by bridging rules (e.g. Carnap's 'meaning postulates') linking some of the terms to measurement and action. We could call that 'symbol-attachment': the role of symbols in an inference mechanism is often prior to reference.

Precocial biological species, competent from birth/hatching clearly refute extreme symbol-grounding theory: foals and chicks don't have time to ground their symbols before using them.

What we really need is *'symbol-attachment'* theory for altricial animals and robots.
See http://www.cs.bham.ac.uk/research/cogaff/talks/#meanings

(Symbol grounding would not explain how explanatory theorising is possible.)

Sometimes the ability to plan is useful

Deliberative mechanisms, possibly using 'attached' but not necessarily 'grounded' symbols with compositional semantics in inference systems, provide the ability to represent unsensed possibilities (e.g. possible actions, possible explanations for what is perceived, possible states of affairs behind closed doors).

One application of that is planning multi-step actions, including nested actions

(unlike 'proto-deliberation', which considers only 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, in a reactive architecture....).

There were also experiments with reactive systems: e.g. simple simulated 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/

Varieties of deliberative mechanisms

What sorts of regions of design space support deliberative capabilities?

Deliberative mechanisms differ in various ways:

- the forms of representations (often 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 kinds of 'compositional semantics' available,
  e.g. Fregean (function application), analogical (picture composition), hybrid forms, etc.
- the number of possibilities that can be represented simultaneously and compared
- 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 oneself or others
  ('meta-semantic' competence linked to meta-management, below).
- the ability to represent abstract entities (numbers, rules, proofs)
- the ability to learn, in various ways, including developing new formalisms, new ontologies, new forms of inference, ...

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

Multi-step planning presupposes discretisation (chunking) of possibilities.

FULLY DELIBERATIVE SYSTEMS

Symbolic AI up to the mid 1980s mainly addressed tasks for which deliberative systems were appropriate.

But only a small subset of deliberative mechanisms was explored, and the processing architectures were not well designed for systems performing most tasks humans can do — e.g. they lacked meta-management.

(Sussman's HACKER – only partially implemented was an exception.)

Some progress was made towards a class of systems with 'fully deliberative' capabilities, including:

- **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, e.g. in neural nets.)
- **The ability to build representations of alternative possibilities, compare them, select one.**

Recently researchers have started adding reflective and meta-management capabilities, using meta-semantic capabilities

E.g. the ability to monitor, detect, categorise, evaluate, plan, debug internal processes including deliberative processes. (See Minsky's draft book *The Emotion Machine*.)

Evolutionary pressures on perceptual and action mechanisms for deliberative agents

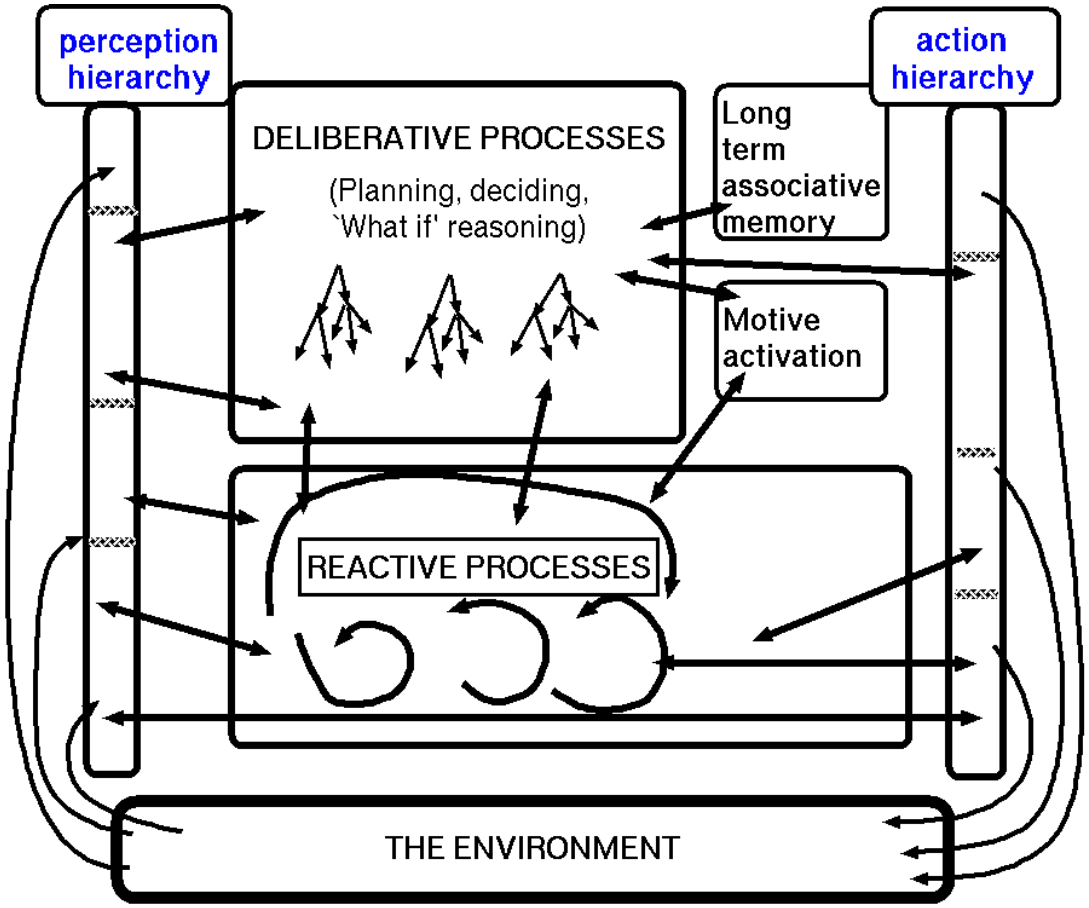
CONJECTURE:

Layered central mechanisms co-evolved with

- **new levels of perceptual abstraction (e.g. perceiving object types, abstract affordances, mental states of others),**
- **new mechanisms supporting high-level motor commands (e.g. "walk to tree", "grasp berry", "express anger".)**

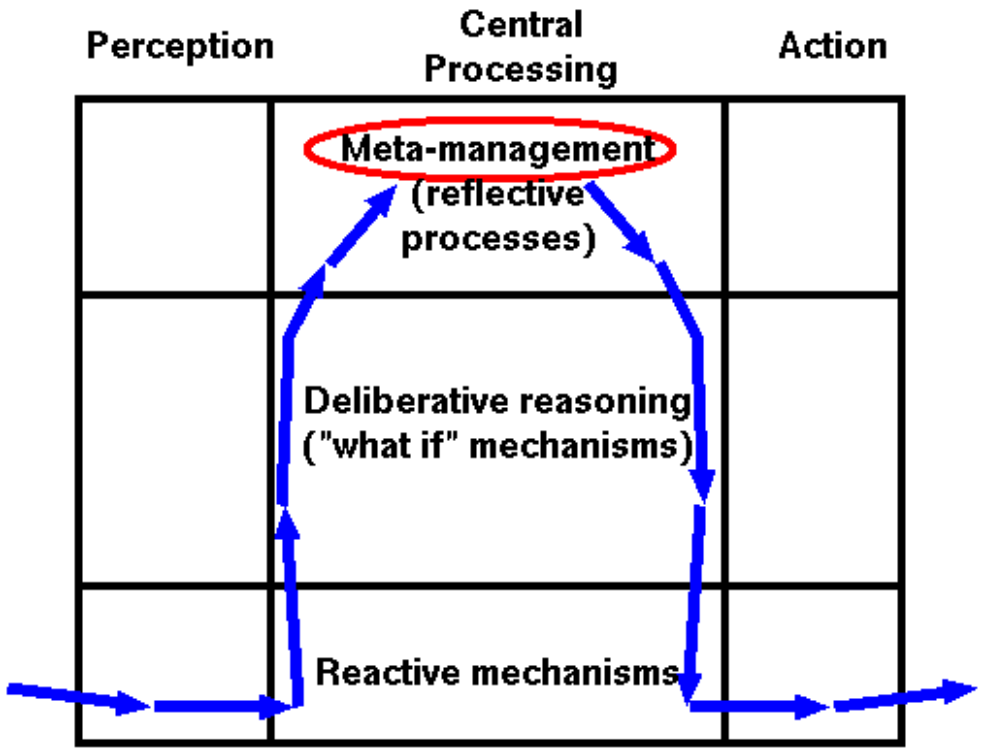
helping to meet requirements for deliberative processes.

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



I call that 'multi-window' perception and action, contrasted with 'Omega' Architectures, which use only 'peephole' perception and action.

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



Compare the greek Capital Omega letter Ω.

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.

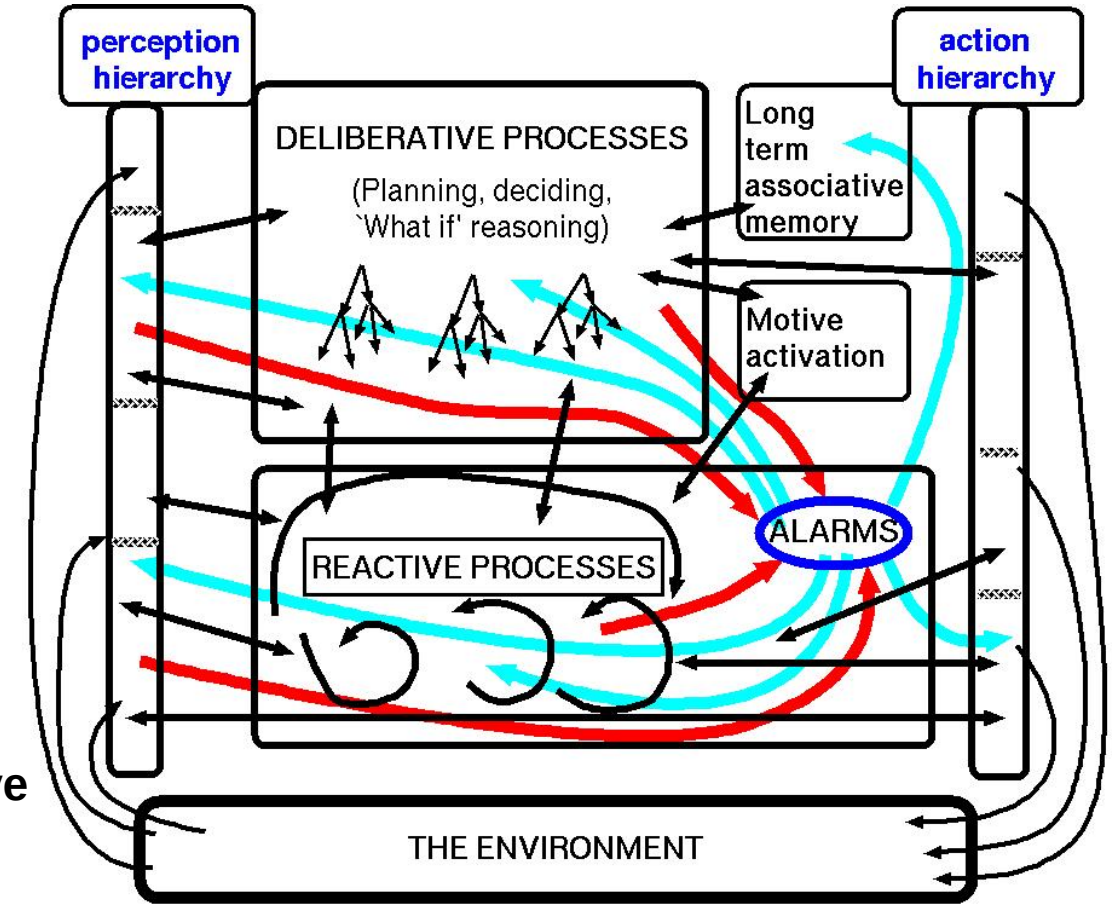
A deliberative system may need an alarm mechanism

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.



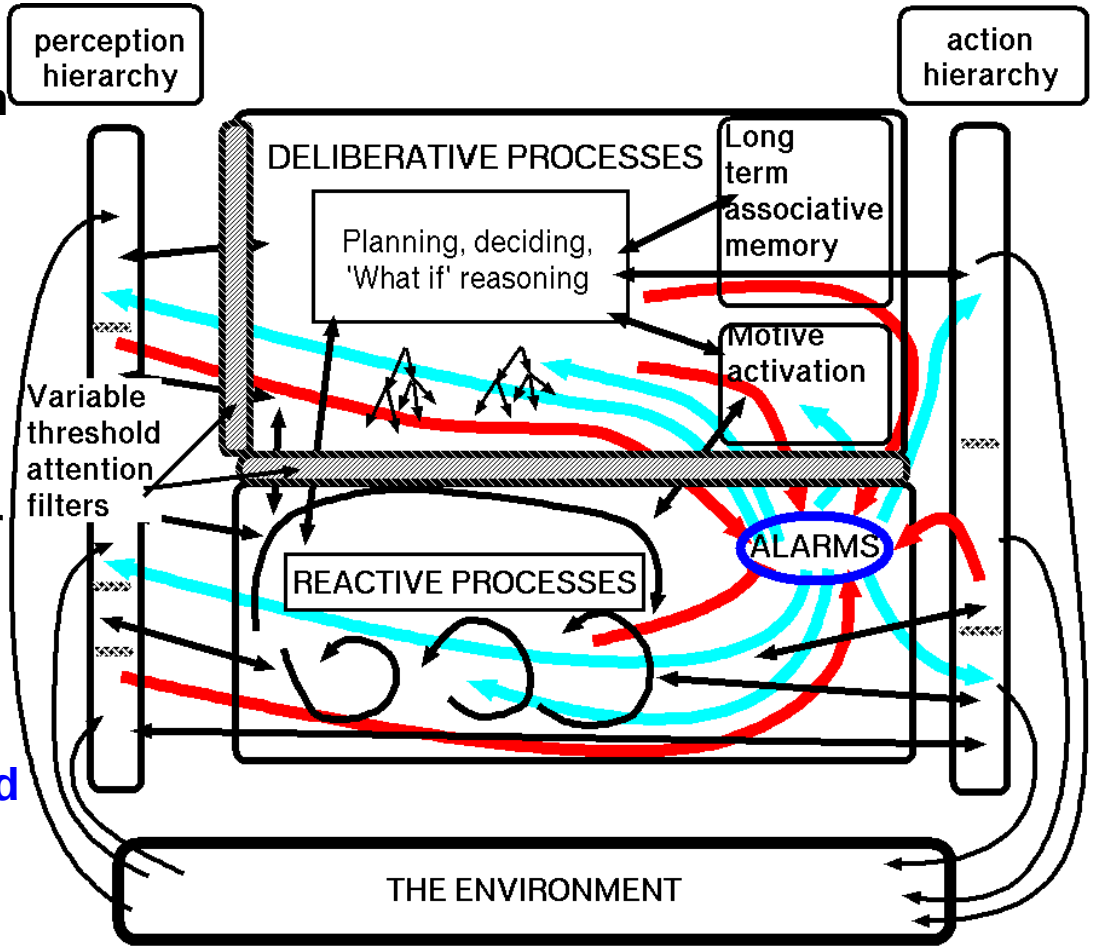
False positives and false negatives can result both from limitations in the learning mechanism and from features of the individual's history: as attested by many aspects of human emotion.

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).

But that's another talk.

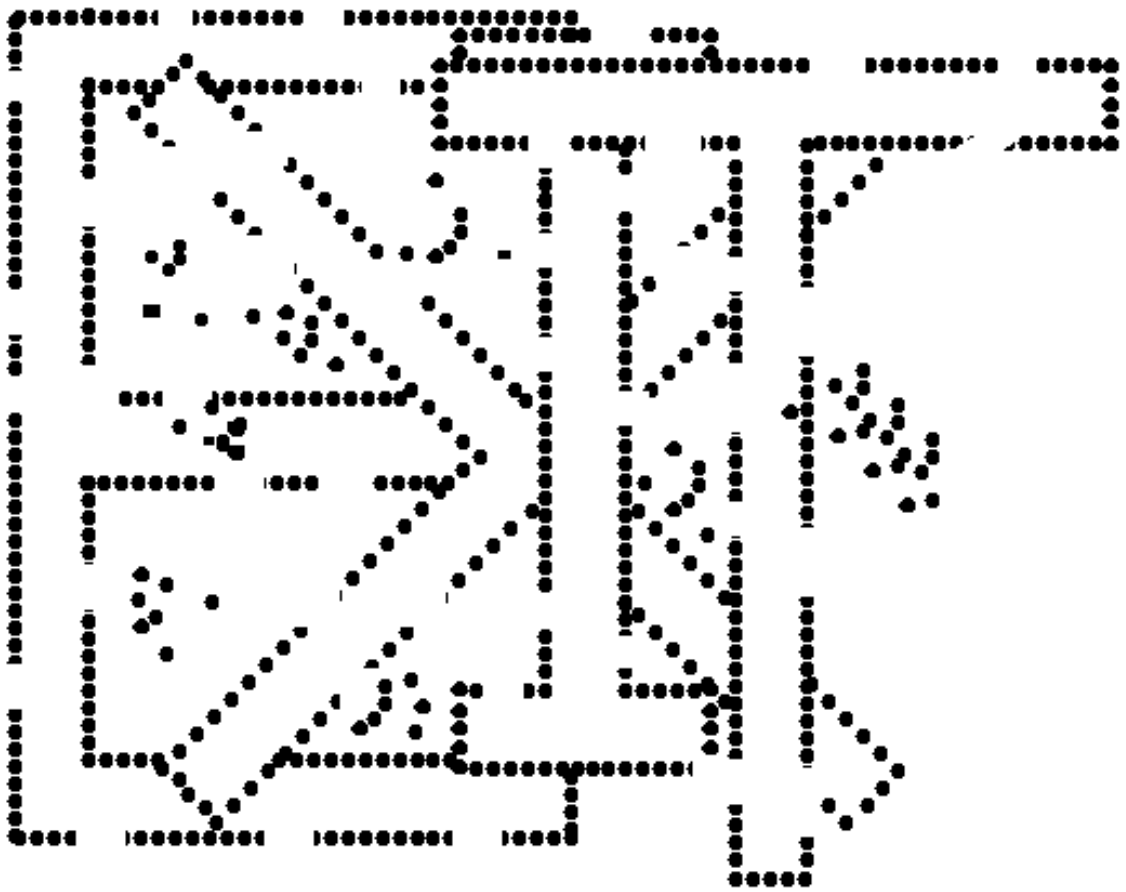
(Compare Freddy II the Edinburgh robot: 1973.)

Vision is an example

Do we know what sort of architecture is required for human-like visual system?

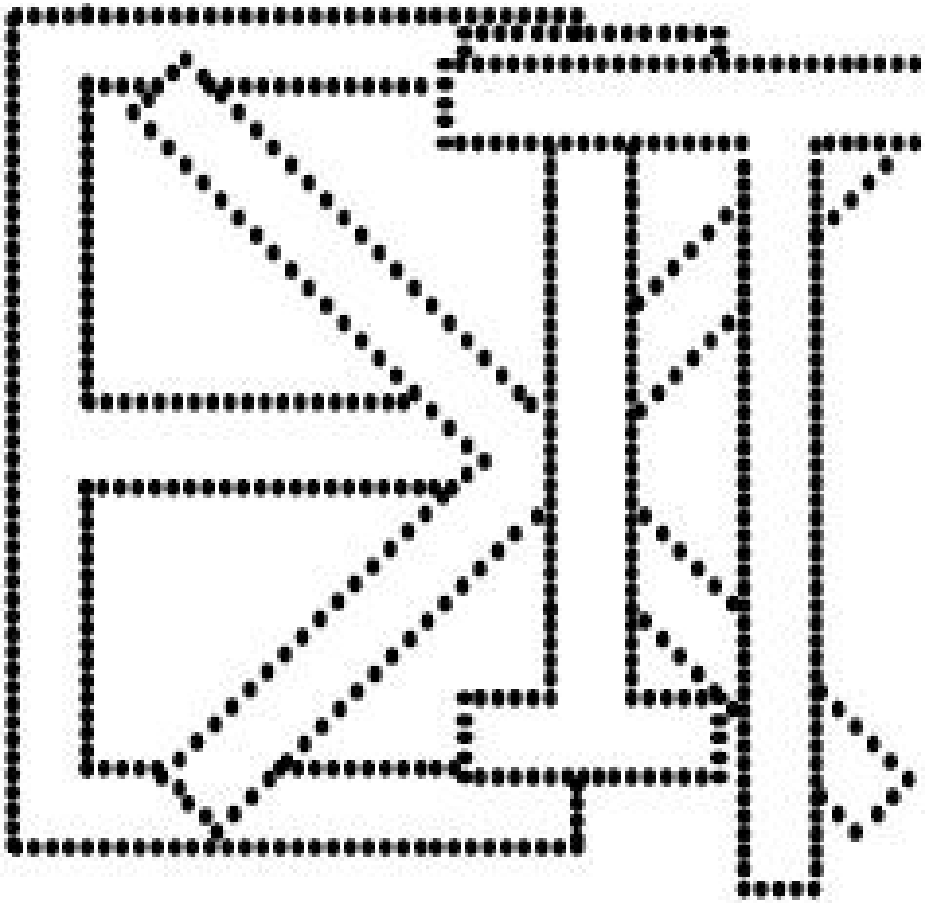
How quickly do you see the word in the next slide??

Try to view it for less than a second.



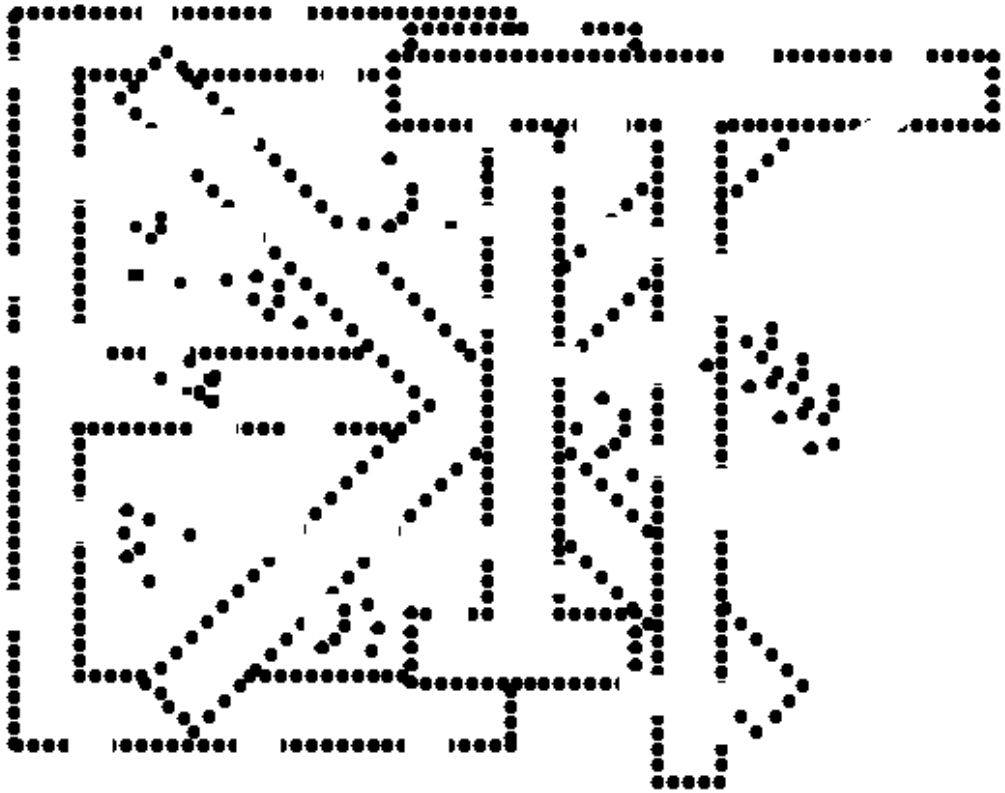
If you did not see a word, try looking back for about two seconds.

Did you see this?



Why do human-like systems need concurrent multi-level perception?

Answer: **In order to cope with rapid recognition of high level structures in complex and messy scenes.**



Despite all the clutter, most people see something familiar.
Some people recognize the whole before they see the parts.

Animal visual systems are not presented with neatly separated images of individual objects, but with cluttered scenes, containing complex objects of many sorts often with some obscuring others.

The objects may be moving, may be hard to see because of poor lighting, or fog, or viewed through shrubs, falling snow, etc.

How do we do it?

Real seeing is often much harder than the tasks most artificial vision systems can perform at present (or tasks presented in vision research laboratories)

Humans (and other animals?) are not always perfect, but they degrade gracefully.

A 30-year old idea may help.

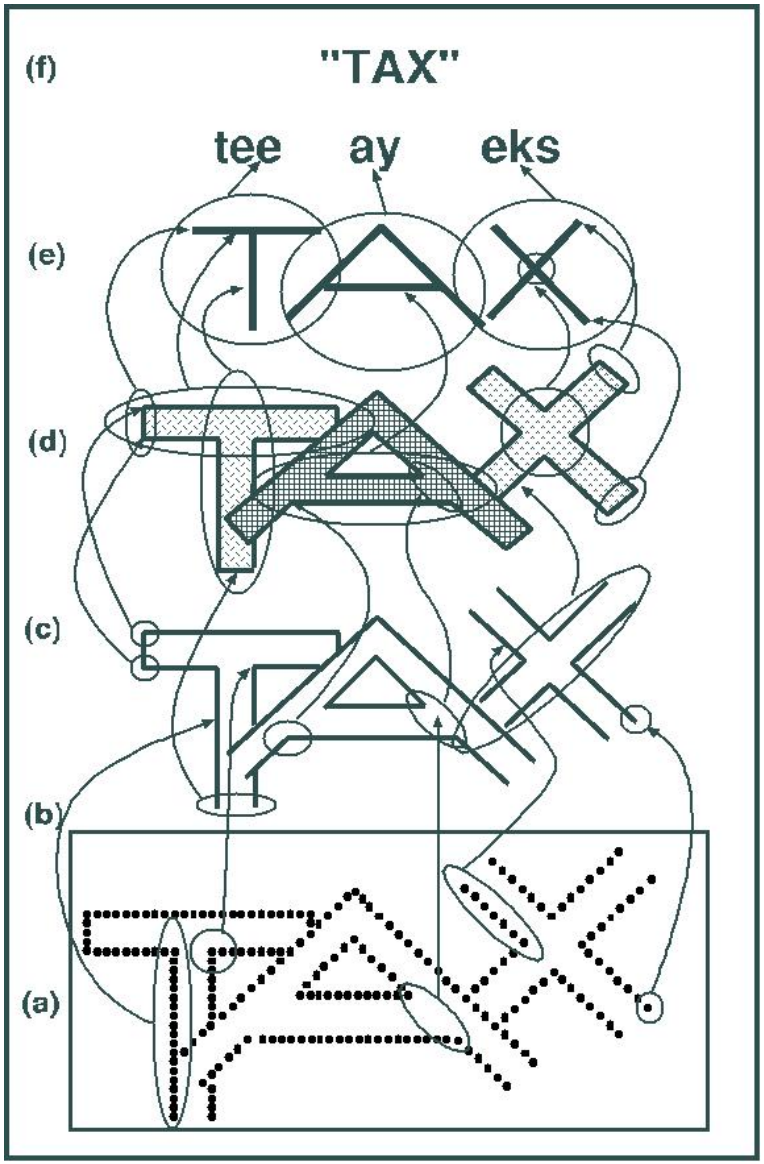
Multiple levels of structure perceived in parallel

Conjecture: Humans process different layers of interpretation in parallel.

Obvious for language. What about vision?

Concurrently processing bottom-up and top-down helps constrain search. There are several ontologies involved, with different classes of structures, and mappings between them.

- At the lowest level the ontology may include dots, dot clusters, relations between dots, relations between clusters. All larger structures are agglomerations of simpler structures.
- Higher levels are more abstract – besides grouping (agglomeration) there is also interpretation, i.e. mapping to a new ontology.
- Concurrent perception at different levels can constrain search dramatically (POPEYE 1978) (This could use a collection of neural nets.)
- Reading text would involve even more layers of abstraction: mapping to morphology, syntax, semantics, world knowledge



From *The Computer Revolution in Philosophy* (1978)
http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html

Multiple visual ontologies, multiple affordances

That was just one example of structural decomposition in a visual architecture.

Altricial animals, like humans, learn about many different domains of structure and many types of affordances, using different sub-ontologies learnt by interacting with the environment (sometimes partly vicariously, e.g. if people are born without arms).

The sub-domains learnt vary across generations as the environment changes, and between cultures, or even between different homes in the same culture.

We don't yet know how many such sub-domains there are in typical human vision.

For more on this see

http://www.cs.bham.ac.uk/~axs/polyflaps

http://www.cs.bham.ac.uk/research/cogaff/sloman-vis-affordances.pdf

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

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*)

- **We call this meta-management (following Luc Beaudoin's 1994 PhD thesis). 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, serving different purposes.**

  (Need to list different purposes.)

So meta-management capabilities evolved

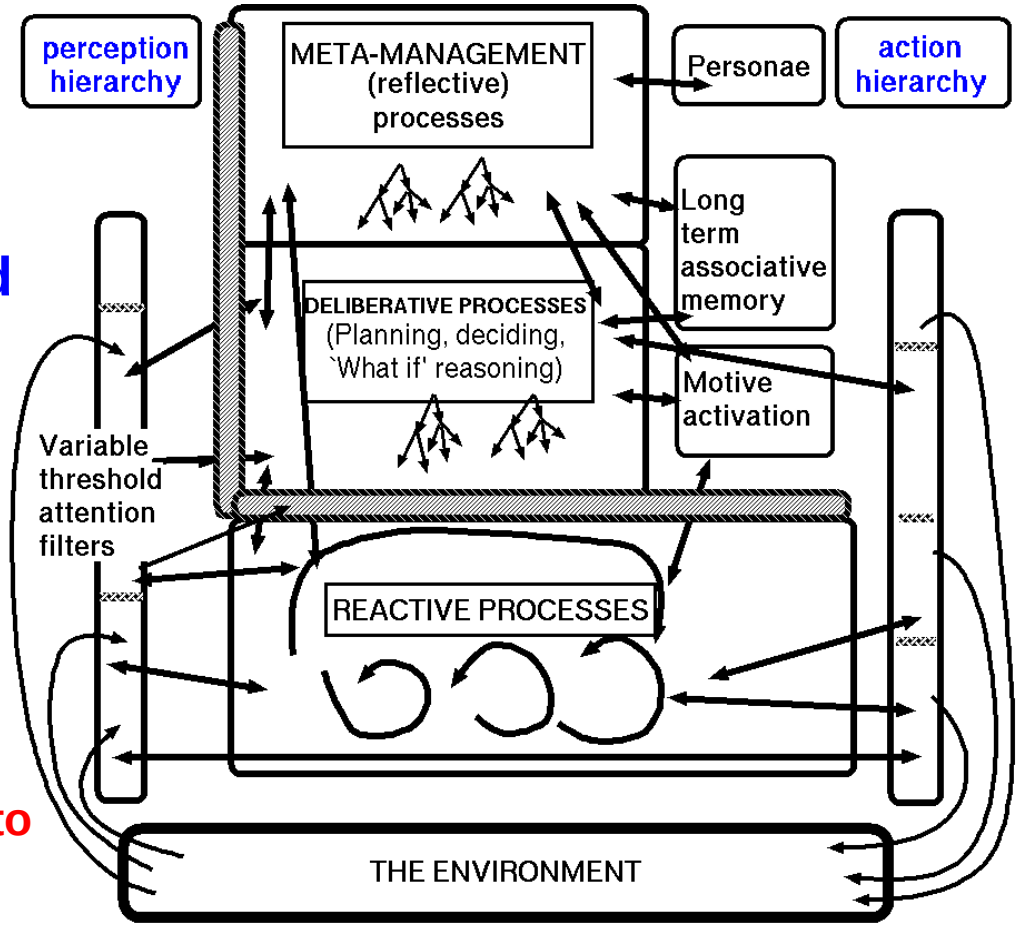
A conjectured generalisation of homeostasis.

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 or damage to meta-management mechanisms can lead to stupid behaviour in AI systems, and in brain-damaged humans.



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

Meta-semantic capabilities used in meta-management also allowed representation of mental states of others, leading to further evolutionary opportunities.

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, or trying to do X.

This is an extension of multi-window perception.

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.

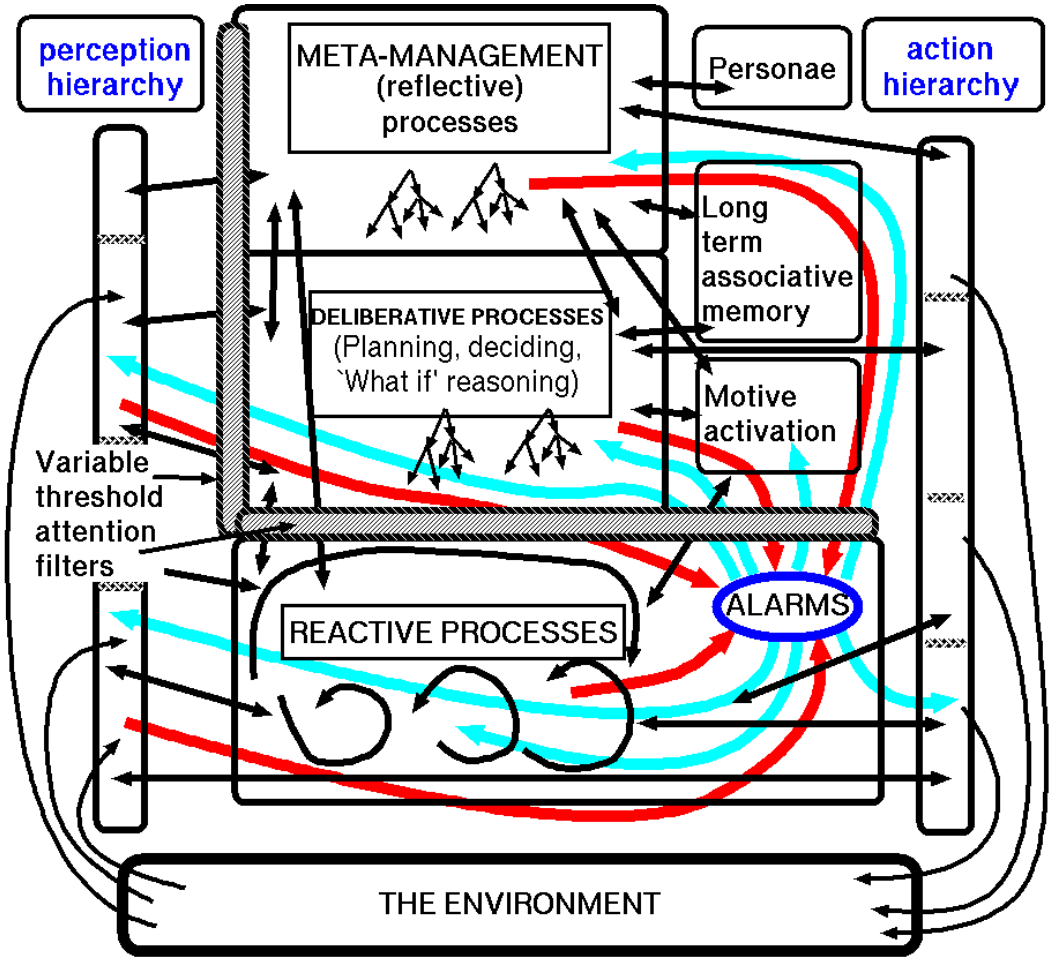
More layers of abstraction in perception and action, and global alarm mechanisms

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.

Where does language fit in?

Clearly language is crucial to humans.

It is part of the process of cultural transmission that accelerates changes in competence, and within individuals it extends cognitive capabilities in many ways (e.g. being able to think about what would have happened yesterday if the weather had not been so bad, and being able to do science and mathematics).

Equally clearly many animals lacking human language have considerable intelligence, shown in hunting, building nests in trees, in social relationships, tool-making etc.

Pre-linguistic human children have many kinds of competence.

CONJECTURE:

In order to understand (and replicate) human languistic competence we need to understand the architectures that suffice for other intelligent species and pre-verbal children, and then see how such architectures might be extended to support linguistic abilities.

It will very likely involve extensions of different kinds in perceptual mechanisms, in all the central processing layers, and in the motor sub-systems.

Mechanisms that proved powerful for development in other altricial species may be crucial for human language learning.

Sloman & Chappell IJCAI05: http://www.cs.bham.ac.uk/research/cogaff/altricial-precocial.pdf

Some Implications

Within this framework we can explain (or predict) many phenomena, some of them part of everyday experience and some discovered by scientists:

- **Several varieties of emotions: at least three distinct types related to the three layers: primary (exclusively reactive), secondary (partly deliberative) and tertiary emotions (including disruption of meta-management) – some shared with other animals, some unique to humans. (For more on this see Cogaff Project papers)**
- **Discovery of different visual pathways, since there are many routes for visual information to be used.**
  (See talk 8 in http://www.cs.bham.ac.uk/~axs/misc/talks/)
- **Many possible types of brain damage and their effects, e.g. frontal-lobe damage interfering with meta-management (Damasio).**
- **Blindsight (damage to some meta-management access routes prevents self-knowledge about intact (reactive?) visual processes.)**

This helps to enrich the analyses of concepts produced by philosophers, scientists and engineers sitting in their arm chairs: for it is very hard to dream up all these examples of kinds of architectures, states, processes if you merely use your own imagination.

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.)
- **Discovery by philosophers of sensory 'qualia'. 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/

The causation problem: Epiphenomenalism

A problem not discussed here is how it is possible for events in virtual machines to have causal powers.

- **It is sometimes argued that since (by hypothesis) virtual machines are fully implemented in physical machines, the only causes really operating are the physical ones.**
- **This leads to the conclusion that virtual machines and their contents are "epiphenomenal", i.e. lacking causal powers.**
- **If correct that would imply that if mental phenomena are all states, processes or events in virtual information processing machines, then mental phenomena (e.g. desires, decisions) have no causal powers.**
- **A similar argument would refute many assumptions of everyday life, e.g. ignorance can cause poverty, poverty can cause crime, etc.**

Dealing with this issue requires a deep analysis of the notion of 'cause', probably the hardest unsolved problem in philosophy.

A sketch of an answer is offered in this Philosophy of AI tutorial presentation: http://www.cs.bham.ac.uk/~axs/ijcai01

See also talks on supervenience and information processing in http://www.cs.bham.ac.uk/~axs/misc/talks/

Many unanswered questions

- About varieties of information
- About forms of representation
- About mechanisms
- About architectures
- About growth and development
- About the variety in the space of possibilities (natural and artificial)

The Future

Some suggestions regarding scenario-based milestones for continuing this research and evaluating progress can be found here

http://www.cs.bham.ac.uk/research/cogaff/gc/targets.html

THERE IS STILL A GREAT DEAL TO BE DONE, BOTH UNDERSTANDING THE PROBLEMS AND UNDERSTANDING POSSIBLE SOLUTIONS.

We all have to learn new ways of thinking.

If we simply continue extending what we have done previously, we shall fail.

I believe that a crucial missing link is understanding mechanisms and forms of representation used in perception of 3-D spatial structure, motion and causal relationships especially as required for manipulating 3-D objects.