TALK FOR COGNIRON WORKSHOP

EURON Conference, Warsaw, 17 Feb 2005

Aaron Sloman

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

School of Computer Science

The University of Birmingham

Towards human-like architectures

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.

Slides available at http://www.cs.bham.ac.uk/research/cogaff/talks/cogniron-slides.pdf

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?

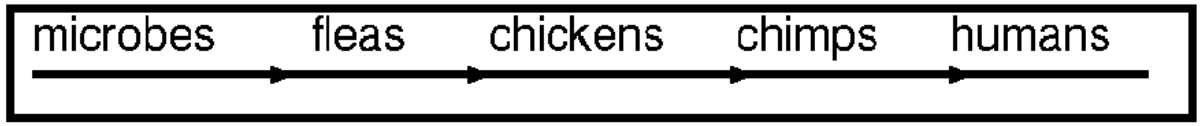
  Answering those questions will help us understand why humans, chimps, lions and crows are (largely) altricial, not precocial like deer, horses, chickens and insects.

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

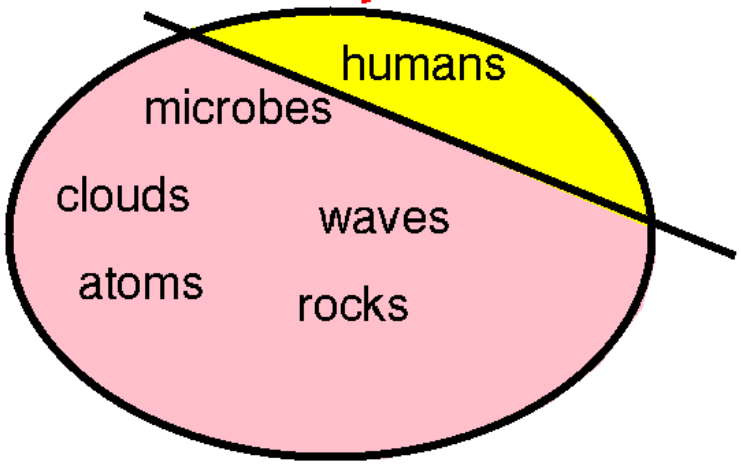
There's No Unique Correct Architecture

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

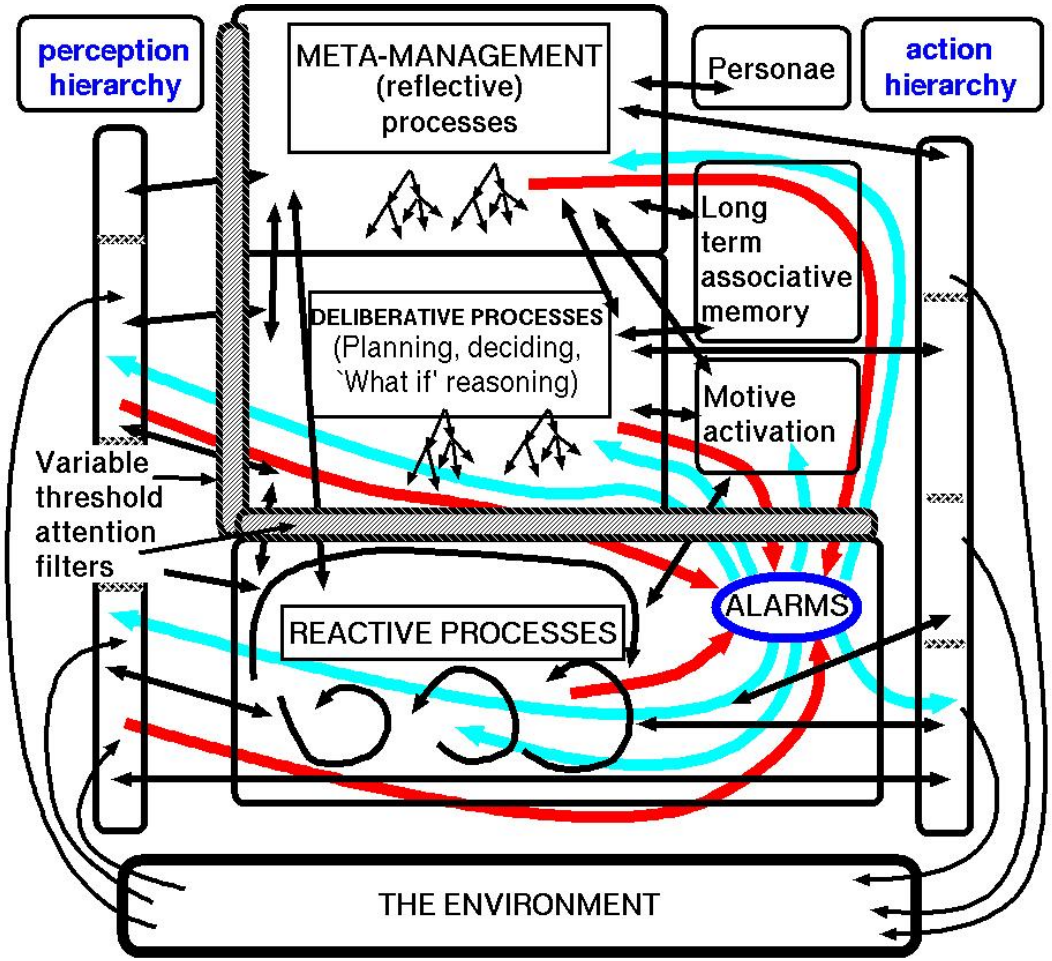


Both 'smooth variation' and a single discontinuity are poor models.

First 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.

Compare: A simple (insect-like) architecture

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

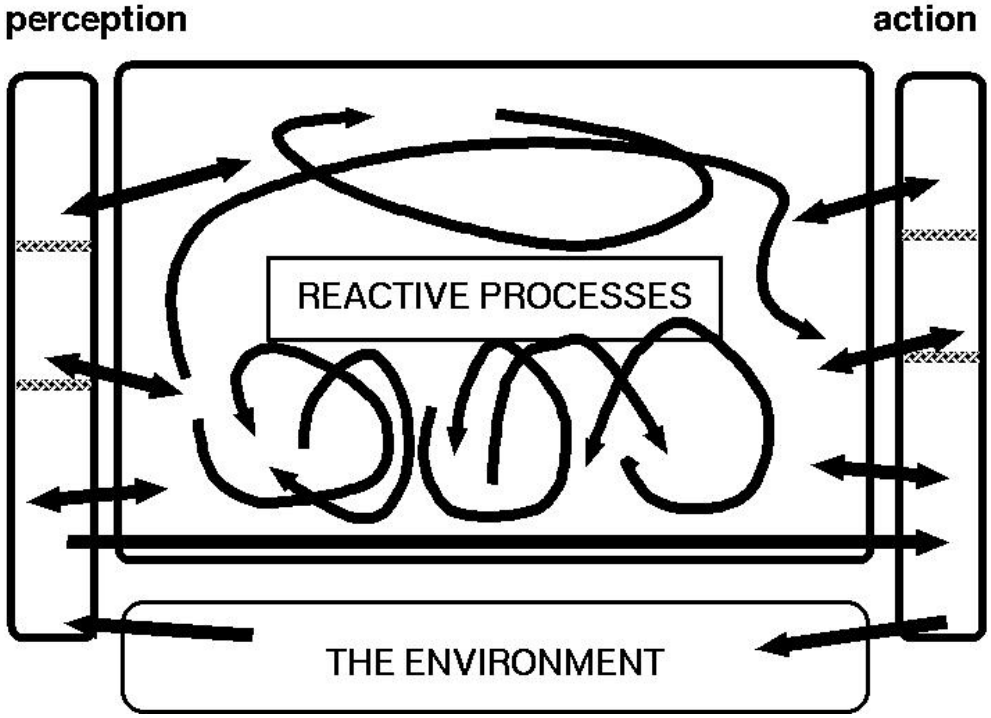
(But see proto-deliberation, later.)

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

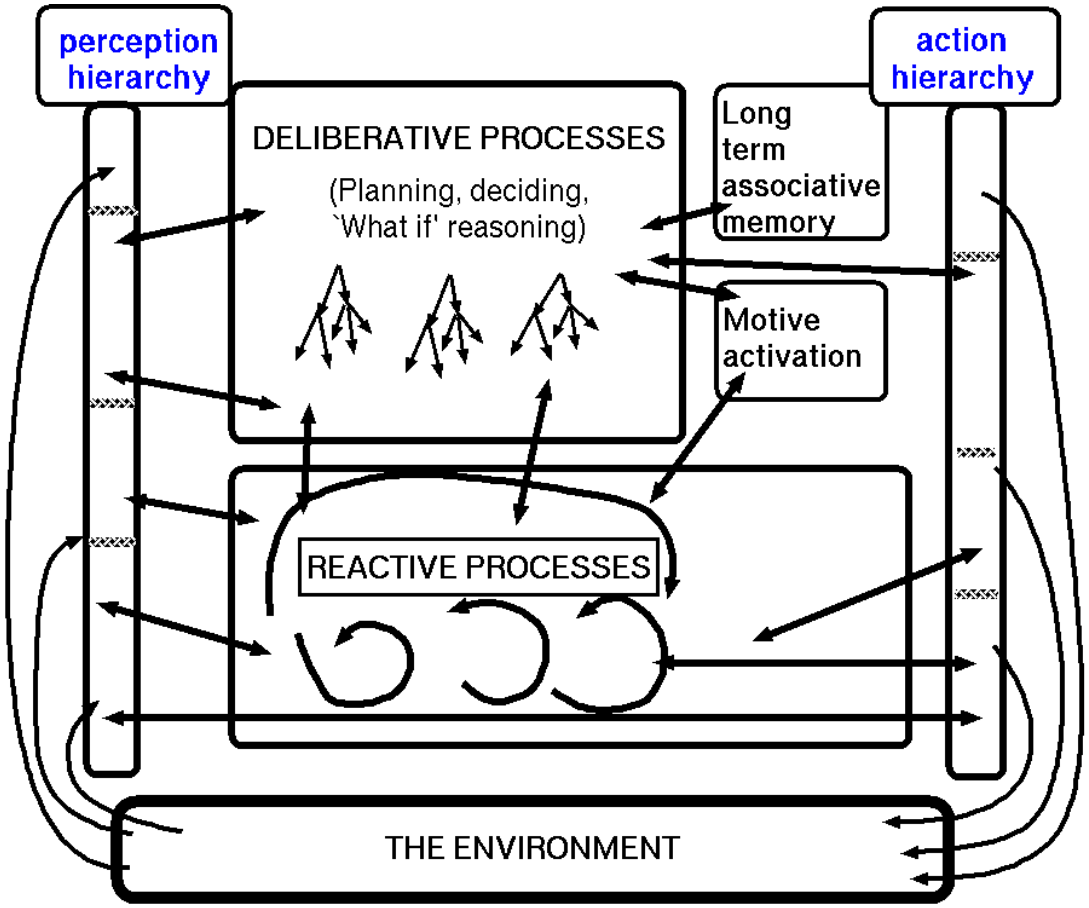
CONJECTURE:

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”.)**

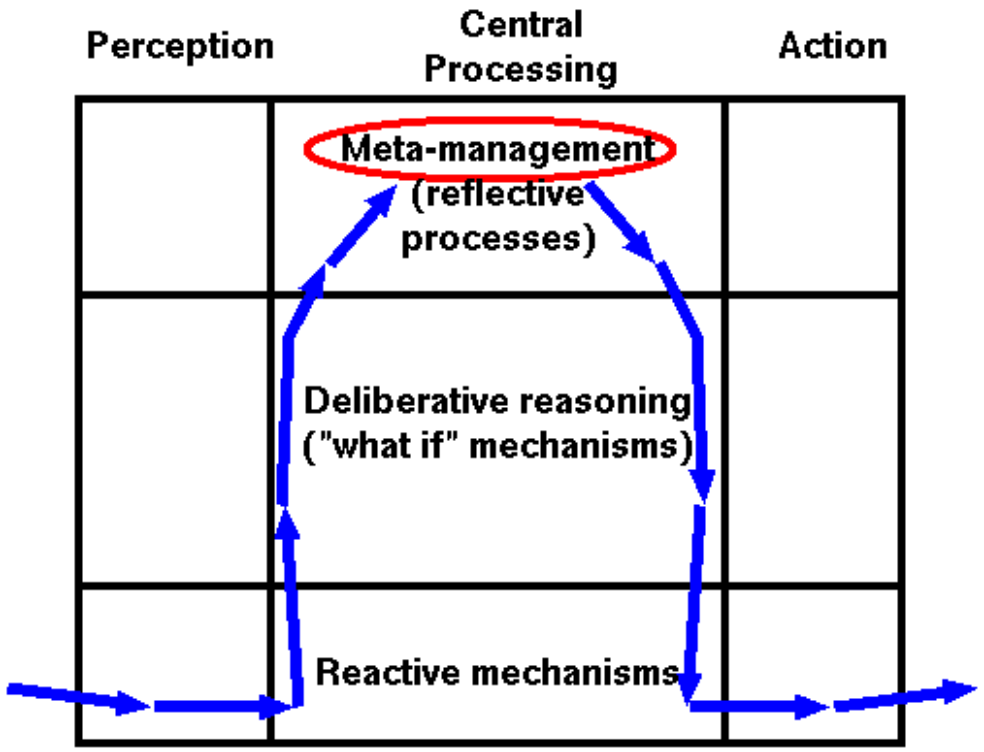
helping to meet requirements for deliberative processes.

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.

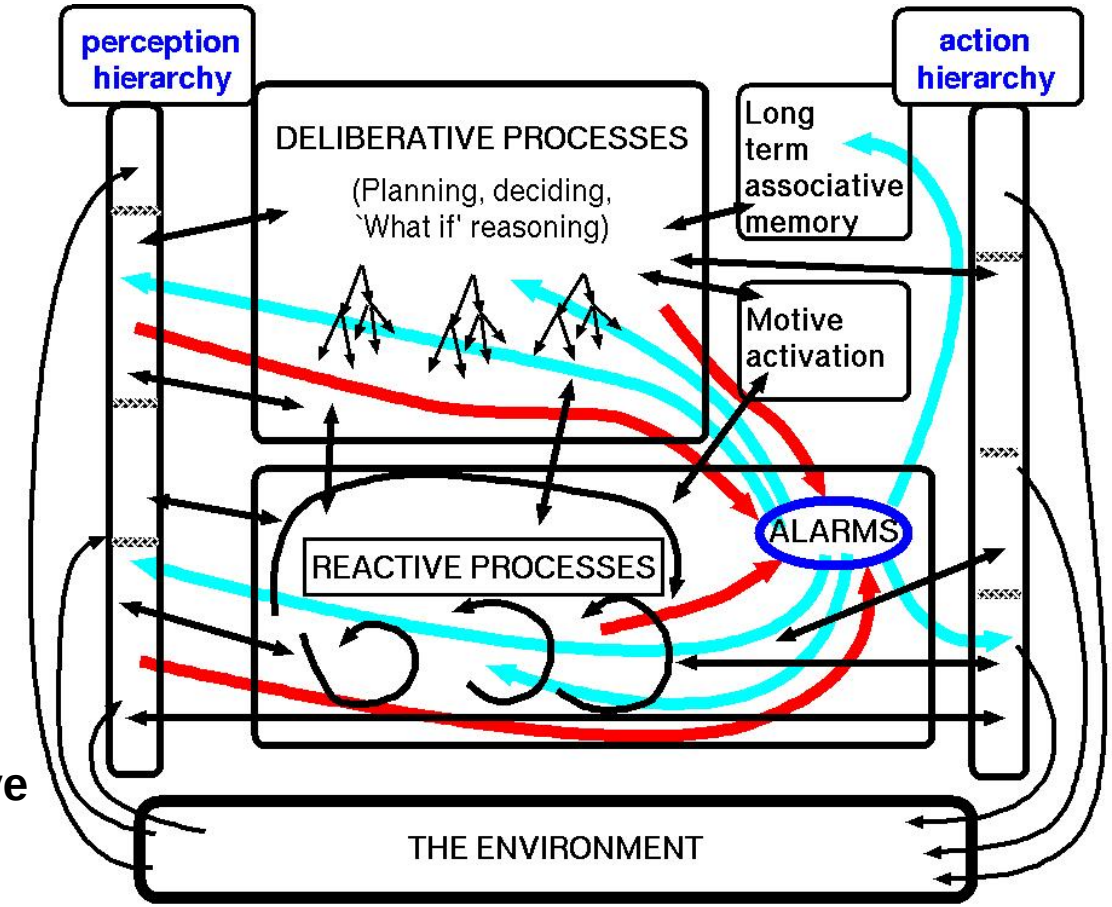
A deliberative system may need an alarm mechanism

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.



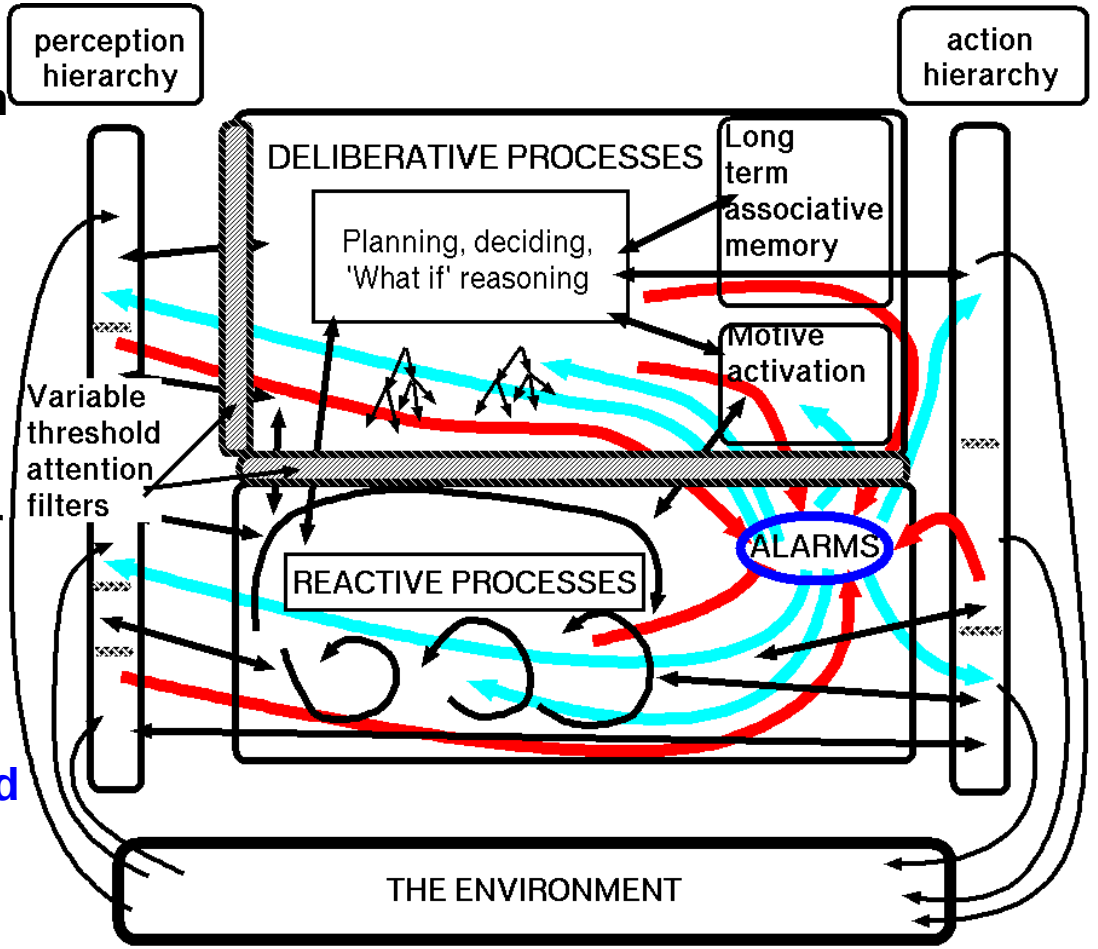
False positives and false negatives can result both from limitations in the learning mechanism and from features of the individual's history: as attested by many aspects of human emotion.

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.

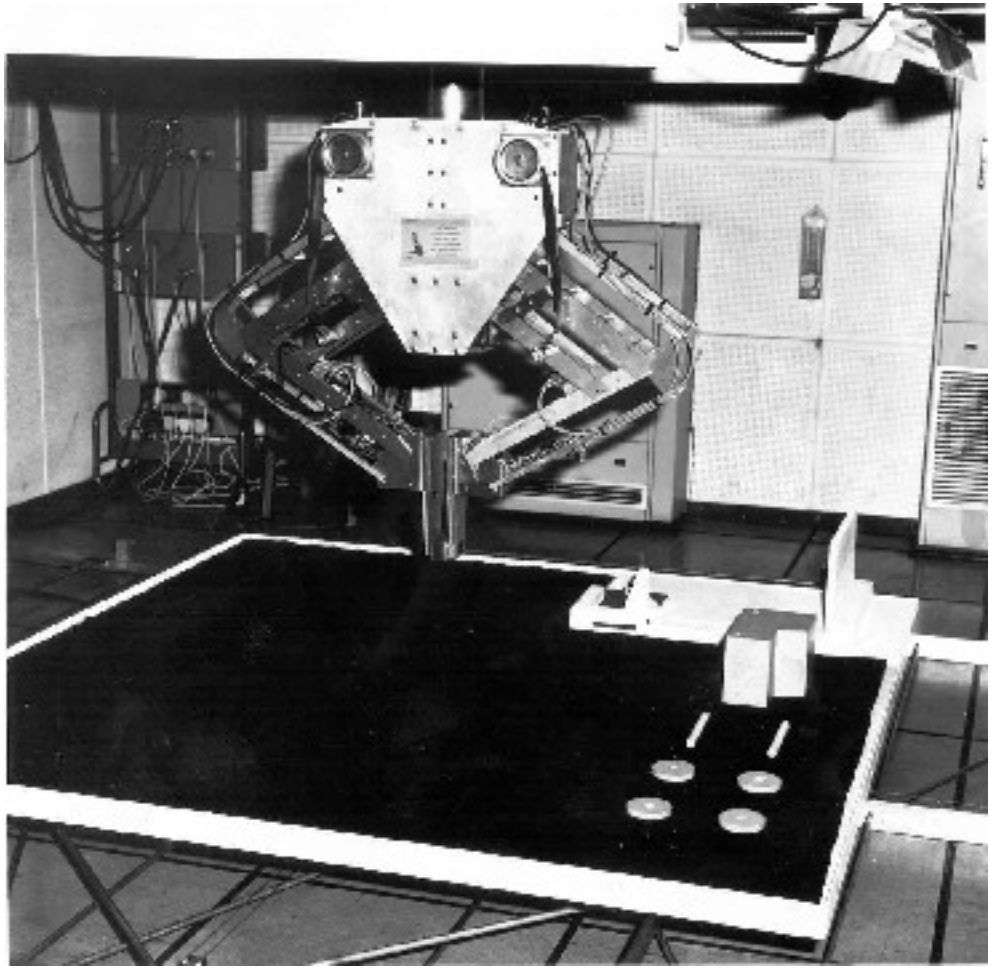
CLAIM:

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.)

FREDDY THE SCOTTISH ROBOT (1973)



http://www.aiai.ed.ac.uk/project/freddy/

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*)

- **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.**

So meta-management capabilities evolved

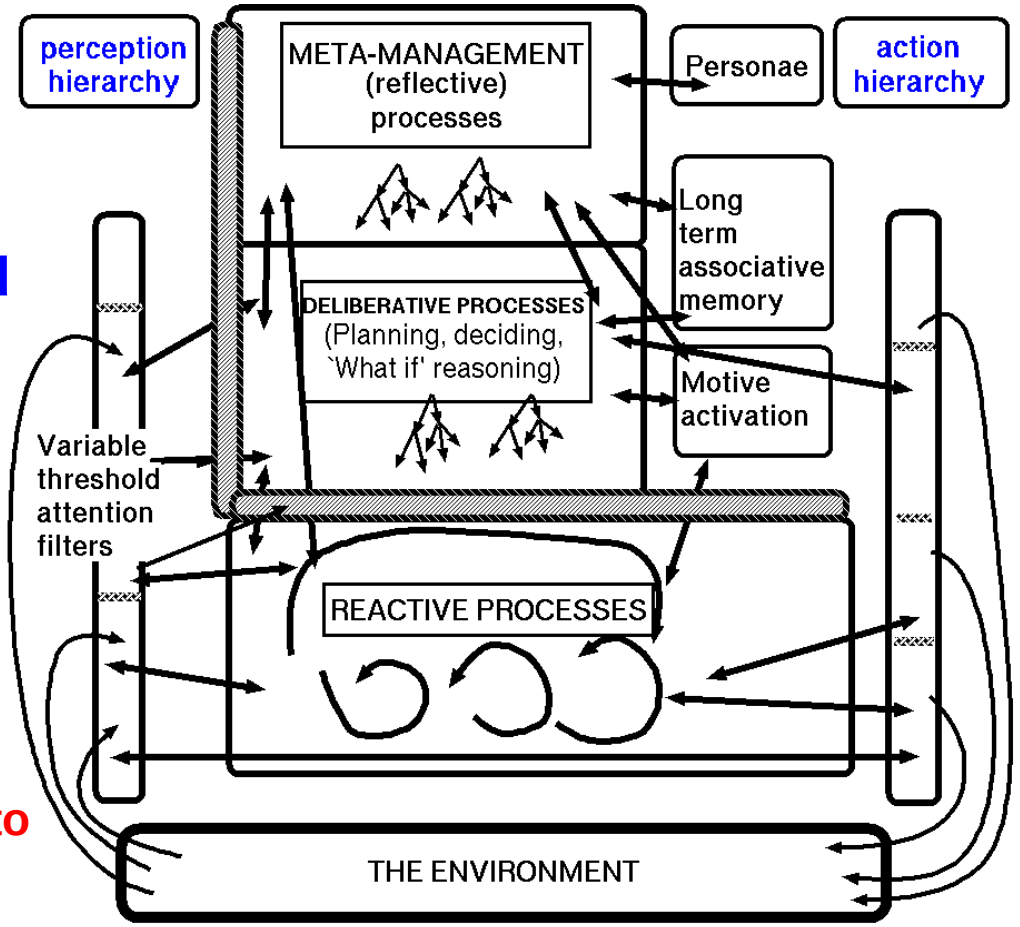
A conjectured generalisation of homeostasis.

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.

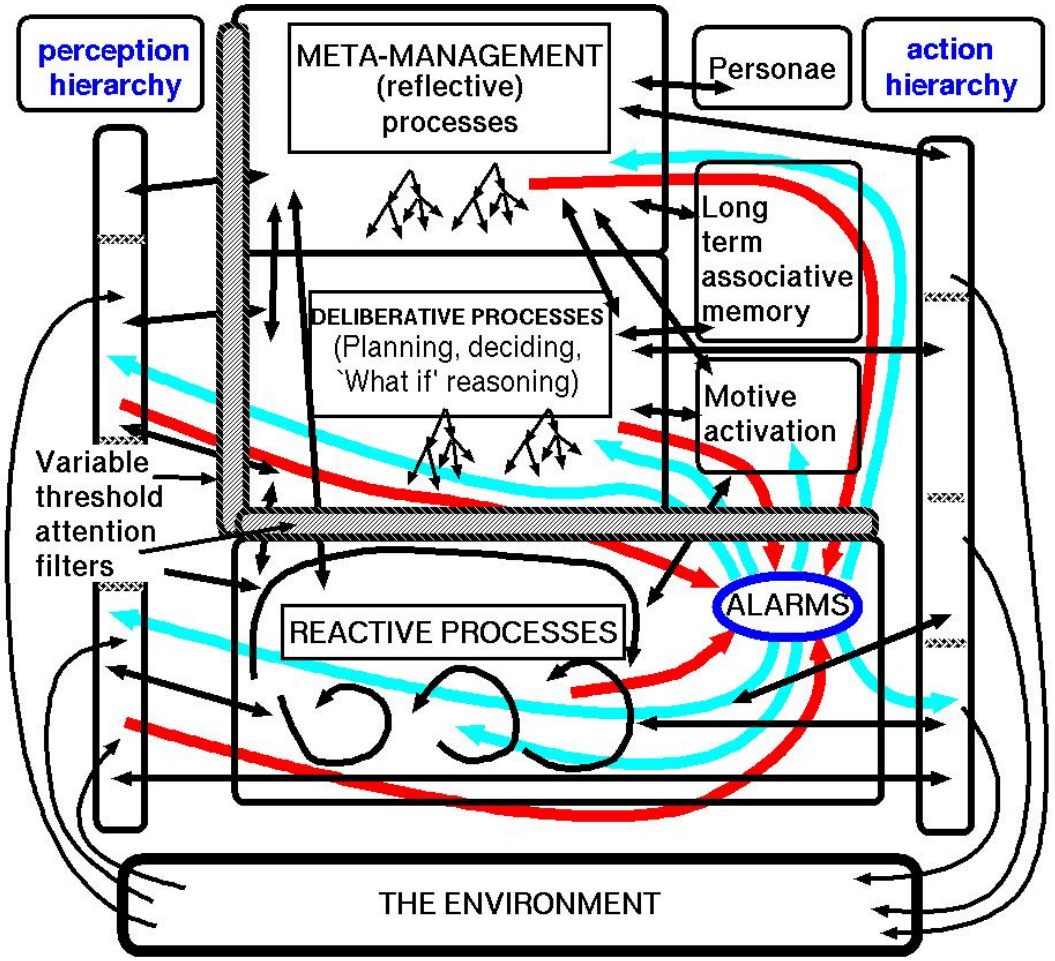
More layers of abstraction in perception and action, and global alarm mechanisms

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.

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 the talk on supervenience and implementation in http://www.cs.bham.ac.uk/~axs/misc/talks/

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.