Architectures for more or less intelligent life

How to turn philosophers of mind into engineers – to help them solve old philosophical problems

Aaron Sloman

School of Computer Science
The University of Birmingham

http://www.cs.bham.ac.uk/~axs/
a.sloman@cs.bham.ac.uk

These slides are available online here
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/#cogarch

Minor re-formatting February 14, 2015

Ideas developed in collaboration with students and colleagues in the

BIRMINGHAM COGNITION AND AFFECT PROJECT
http://www.cs.bham.ac.uk/research/cogaff/

AND THE INTELLIGENT ROBOTICS GROUP
http://www.cs.bham.ac.uk/research/groupings/robotics_and_cognitive_architectures/

[Apologies: these slides were composed a long time ago, before I had learnt to use Latex properly.]

OVERVIEW: How to integrate philosophy of mind

- Consider whole architectures

  (Not just vision, belief, desire, inferring emotion, being rational, ...)

- Consider different species

  (microbes, insects, birds, elephants, ... not just humans)

- Consider developmental and evolutionary trajectories

  (what changes occurred in our evolutionary precursors, and between: infants, toddlers, teenagers, professors,..., and also social/cultural progressions)

- Consider individual differences

  (autism, down syndrome, blind, deaf, brain damaged, limbless; pianist, athlete...)

- Consider past and future artefacts (e.g. robots, not just biological systems)
- Consider design requirements (niches, and how they can vary and have changed)
- Consider design possibilities (from microbes to ecosystems and human societies)
- Combine multiple disciplines (including philosophy)

  (Switch modes of thinking often!)

- Acknowledge conceptual confusions in some philosophical problems

  (They don't necessarily address sensible questions – e.g. "Where is the universe?")

The importance of architectures

When trying to understand a complex system with many functional components we need to know how the components are related, and which ones interact with which others, or with the environment and how.

In other words: we need to try to understand the architecture in which all the components are combined.

So we need good ways of thinking about architectures – how they can vary and what the consequences of differences are.

The Cognition and Affect project has provided a partial analysis of the space of possible architectures, but there's a lot more work to be done:

`http://www.cs.bham.ac.uk/research/projects/cogaff/#overview`

In contrast, many researchers try to propose just one architecture as the right one.

But there is no reason to believe that the information processing architectures of a new-born infant and a typical human adult are the same.

There may also be architectural differences between different adults.

Instead of thinking in terms of how things work, most philosophers tend to think in terms of states of the whole person.

E.g. X believes P if X .... (like trying to explain what water is in terms of how it looks, feels and tastes to us, rather than its chemical composition).

Architectures need not be physical architectures
We are just beginning to understand virtual machine (VM) architectures

A philosophically significant development in the last half century has been the creation of increasingly varied and sophisticated virtual machines, including running interpreters, games (e.g. chess), operating systems and their many components, and networked systems,

“Virtual” does not mean “unreal”, or “imaginary” or “lacking in causal powers”.

Virtual machines in computers are as real as poverty, economic inflation, and other abstract processes that impact on our lives.

All of these have causal powers, and are therefore not “epiphenomenal”

They are “emergent” phenomena with causal powers.

But nothing spooky! Engineers design some of them.

Philosophical significance:

- This extends our ideas about kinds of causes and effects, providing new ways to think of mental processes as being able to change the world.
- Understanding the benefits of VMs for engineering design purposes (including self monitoring and self modification) can give us new ideas about how and why minds based on virtual machinery might have been produced not only by engineers, but also by biological evolution – with similar benefits.

Ask the Ghost of Gilbert Ryle

In 1949 Ryle published The Concept of Mind, criticising the idea of “the ghost in the machine”.

However, some of his own work, e.g. Chapter 8 on imagination, supports (reluctantly?) sorts of ghosts (virtual machinery, though he does not use that label) within physical machines (biological bodies).

Every intelligent ghost must contain a machine

What sort of machine is required for intelligent behaviour?

An information-processing machine.

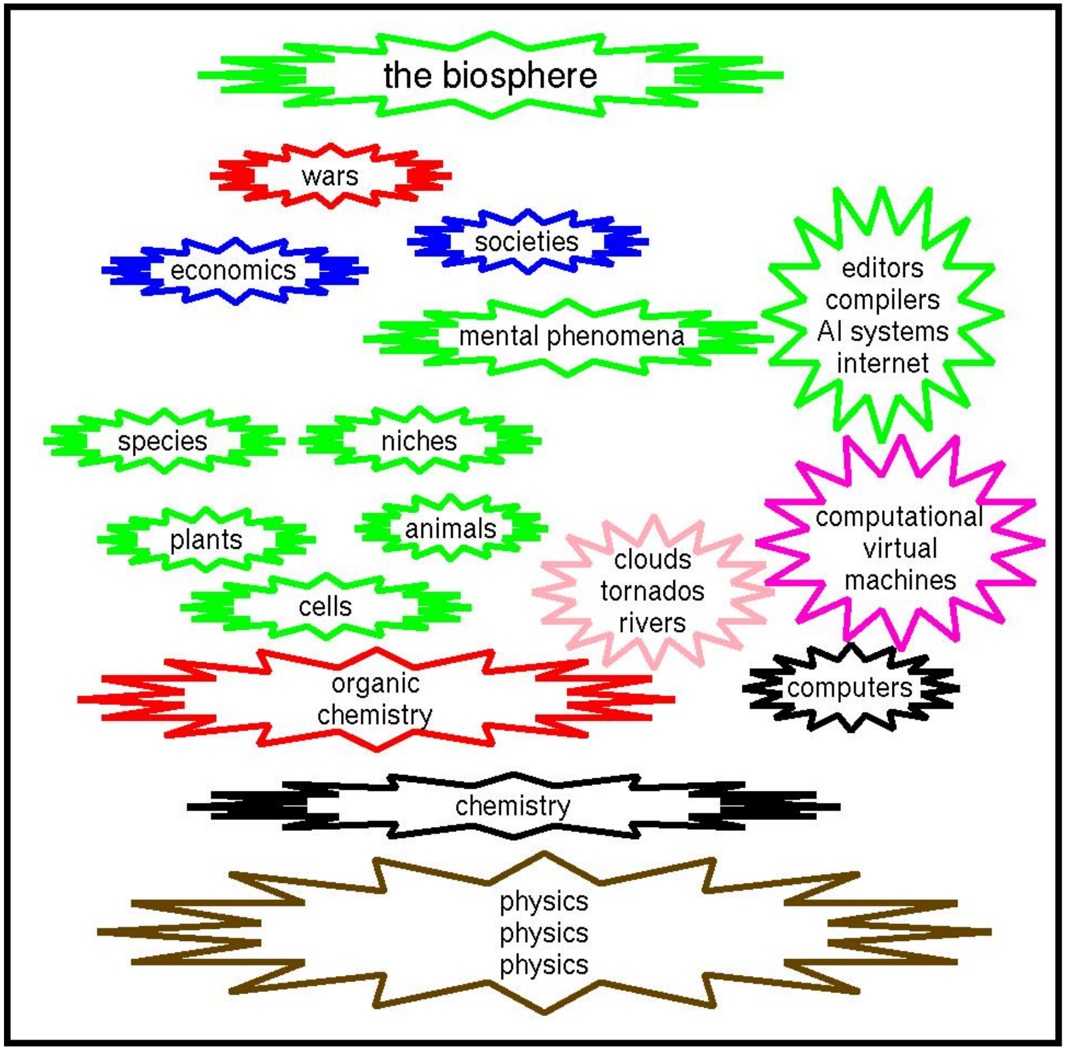
Not necessarily a physical machine

since many of the things going on in such a machine, e.g. wanting, inferring, supposing, deciding, imagining, intending, wondering,... are not describable in the language of physics,

But the information-processing machine must be **implemented** in physical mechanisms if it is to DO anything.

(Why is this necessary?)

Emergent virtual machines are everywhere



How many levels of physics will we know about in 500 years time?
Different levels involve different ontologies.

Philosophy and Engineering

A common comparison:

MIND $\Longleftrightarrow$ BRAIN

VIRTUAL MACHINE $\Longleftrightarrow$ PHYSICAL MACHINE

The first relation $\Longleftrightarrow$ is often referred to as "supervenience", the second as "implementation", or "realisation", or "support", understood by software engineers.

Philosophers usually discuss supervenience in ignorance of what software engineers know or do.

And most engineers design and build their systems without noticing or thinking about its philosophical relevance.

However, the situation is changing slowly – especially in AI,

since AI researchers often wish to make philosophically significant claims about what they are doing (e.g. enabling machines to see, intend, have emotions, discover concepts, be creative, communicate).

Note that what software engineers, or computer systems engineers, actually do is very complex, and hard to make precise and clear – even professionals can mis-describe it.

For example, many don't understand the importance of parallel streams of causal influence within the virtual machinery interacting with the underlying physical machinery and the environment.

Different sorts of supervenience

- Pattern supervenience:
  being arranged in regular columns supervenes on being arranged in regular rows.
  (This can hold in both directions.)

- Agglomerative/additive supervenience:
  a complex object weighing 3Kg supervenes additively on the weights of all the parts.

- Property supervenience:
  one property (e.g. being symmetrical) supervenes on another (e.g. having certain measurements).

- Mechanism supervenience:
  one ontology, including structures, processes and causal interactions, supervenes on another.

  Mechanism supervenience, with one ontology supervening on another, is something we have only recently come to understand, through being forced to build increasingly complex examples, in order to get machines, including networks of machines, to perform tasks that no previous technology could do.

  But we still have much to learn.

We understand only a tiny subset of the space of possible virtual machine architectures.

Different VM architectures are required for different engineering applications

e.g. spelling checkers, chess machines, operating systems, distributed databases, mobile phone networks, control system for a chemical plant, ...

and also for minds of different sorts

(e.g. adult human minds, infant human minds, chimpanzee minds, rat minds, bat minds, flea minds, damaged or diseased minds ....).

We need to place the study of (normal, adult) human mental architectures in the broader context of

THE SPACE OF *possible* MINDS

I.e. minds with different architectures that meet different sets of requirements, or fit different niches.

Deep understanding will not come from studying ONE case a typical adult human mind! Which is all most philosophers do.

Let’s look at neighbourhoods and trade-offs

- in design space
- in niche space

Let’s analyse:

- trajectories of different types through these spaces,

  in evolution, in individual development, in learning, in cultural change, in repairing, bug-fixing ...

- interactions between the trajectories, e.g. the many feedback loops in co-evolution.

- architectures, not only for individuals, but also for components and larger structures:

  e.g. families, teams, pairs fighting, economic systems, eco-systems.

No portion of either space will be fully understood without the context of the remainder.

Only trivial things can be taught or studied without generating confusion that is later cleared or reduced, by providing more information, especially many more examples.

IS EVOLUTION A DESIGNER?

Yes insofar as it produces designs:

- Partly implicitly, by producing instances of those designs
- Partly by producing re-usable specifications for designs using powerful formalisms:

  (genetic code or codes, which we only partly understand)
- Also in using information in the process:

  information that is mostly implicit and scattered among all the co-existing, co-evolving species: information about varieties of environments, and what does and does not work in various environments.

But evolution is a “reactive” system, not a “deliberative” designer

(it does not build representations of various possible futures and compare them).

It also lacks “meta-management”

(it cannot inspect and evaluate its own information processing capabilities to decide what to fix, or modify).

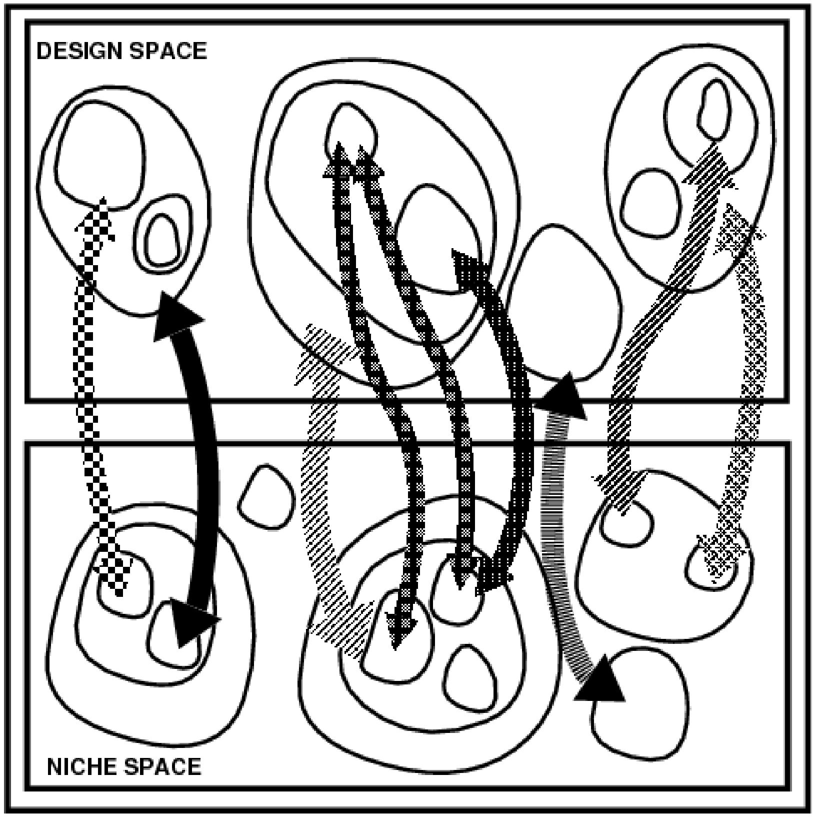
A possible exception:

evolution can (in effect) use the cognitive abilities of “intelligent” informed individuals. (Evolution uses many of its products as helpers, enhancers. e.g. in mate selection.)

Evolution produces *changing niches* as well as *changing designs*

The whole process has many complex, constantly changing, concurrently active feedback loops between niches and designs.

DESIGN SPACE AND NICHE SPACE



Relations between designs and requirements (niches) are not just “fitness functions”.
They are multi-dimensional relationships. (Like evaluations in consumer reports.)

A design can be related to many possible niches and *vice versa*.
(Multiple mappings not shown in the above diagram.)

Designs can have different levels of abstraction.

There are different sorts of trajectories through design space and niche space

i-trajectory: possible for an individual organism or machine, via development, adaptation and learning processes (of many types): egg to chicken, acorn to oak tree, etc.

e-trajectory: possible for a sequence of designs evolving through natural or artificial evolution. Requires multiple re-starts in slightly different locations.

s-trajectory: possible for social systems with multiple communicating individuals. (Can be viewed as a type of i-trajectory.)

c-trajectory: trajectory made possible by the use of cognitive capabilities of individuals, e.g. mate selection or differential parental caring for young of different capabilities.

r-trajectory: possible for a system being repaired or built by an external designer whose actions turn non-functioning part-built systems into functioning wholes.

All but r-trajectories are constrained by the constant requirement for “viable” systems.

R-trajectories used not to be possible for living things, but development of increasingly sophisticated prosthetic devices is changing that: e.g. artificial hands, kidneys, pacemakers...?

In all types, “search spaces” can be astronomical, or worse.

Trajectories depicted

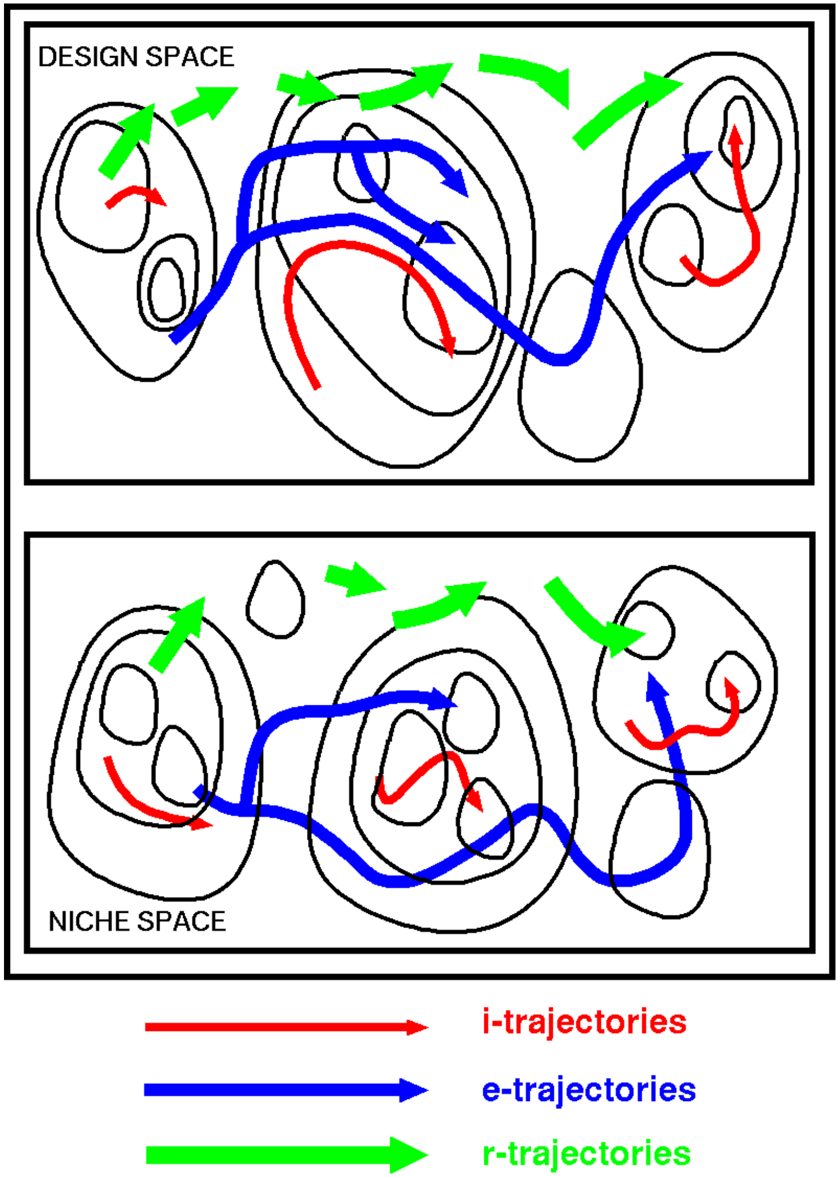
An external “repairer” can push something through a discontinuous “r-trajectory”

(going through intermediate states that are not viable.)

Biological evolution is inherently discontinuous not continuous, though many of the discontinuous changes are small.

Chemical changes involving destruction and construction of complex molecules are inherently discontinuous: A molecule can contain only a discrete number of atoms of carbon: the number cannot change continuously.

Biological building blocks are molecules.



Biological evolution uses interacting trajectories

Biological evolution requires parallel evolution of different sorts of organisms, including some near the peak of a food pyramid.

Multiple interacting e-trajectories,

later using i-trajectories,

then s-trajectories and c-trajectories,

and now also human-designed r-trajectories

(prosthetics, ... genetic engineering, ... ?)

Under what conditions does the (expensive) transition to deliberative capabilities pay off, compared with other design options? (Explained in (Sloman, 2006))

Are those conditions very rare?

The evolution of cognitive mechanisms can produce “c-trajectories”, which use the *cognitive* abilities of individuals to modify e-trajectories.

Many questions: e.g. why are there so few “intelligent” species or individuals.

Whether we count: species, individuals or biomass.

Part of the answer: having intelligence of the sort that requires big brains implies being near the top of a large food pyramid.

Many everyday concepts are unfit for scientific use

‘Mind’, ‘emotion’, ‘consciousness’, ‘intelligence’, ‘representation’, ...
are all too ill defined.

We can use architecture-based concepts to refine and extend such concepts.

E.g. Understanding the architecture of a complex system enables us to work out many of the kinds of states and transitions that can occur, and why.
So we can use unobservable features to distinguish cases that look the same from their behaviour, ... and group together similar internal states and processes that look different from outside.

Can you think of examples (a) in computing systems (b) in human minds?

Compare the way advances in science helped to refine and extend our concepts of animal, mammal, fish (e.g. excluding whales), water, salt, iron, gold, and many other chemical elements and compounds.

Compare physics: the architecture of matter

But beware: there’s not just one architecture for mind

For examples of confusions about emotions see
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cafe04`

See also (Sloman, 2002)

EVOLUTION OF MIND

Different mental concepts are applicable in different architectures

- An architecture supports a collection of possible states, processes, causal interactions, and forms of development:
- Different collections are possible in different architectures. (Why?)
- If mental concepts are architecture-based then we can't apply the same concepts to organisms with different information-processing architectures
- Compare:
  - A fly that is "conscious" of my rapidly approaching hand
  - An adult human "conscious" of a rapidly approaching mugger's fist
  - Can a goldfish long for its mother – if not why not?
- Do not expect to be able to use your concepts to make sense of questions like

  "What is it like" to be a fly, a bat a new born baby ?
- So, when we understand how information-processing architectures change during evolution we'll be better able to understand how new mental states and processes become possible.

Are Babies Designed to Deceive?

Doting parents, and experimental developmental psychologists try to work out what’s going on in the minds of new-born infants: but perhaps the concepts they use to formulate questions are inapplicable to systems with such under-developed architectures.

Perhaps evolution designed human babies with the ability to fool parents into treating them as humans

so that they get treated in a way that helps them build their human information-processing architecture?

Architectural diversity

Even apparently similar animals may have surprisingly different information processing virtual machine architectures

- Some types of bird can remember individual locations of many nuts they have hidden and which ones each has eaten.

  This requires a web of spatial (including probably both metrical and topological) relationships to be perceived and the information stored – then changed as individual nuts are consumed.

- Other species cannot. How they perceive their environment will be importantly different.

- There are broad distinctions between the developmental patterns of different species.
  - Precocial species are born or hatched ready to feed, walk, swim, run, etc.

    (e.g. chickens, deer, horses...)
    `http://www.bbc.co.uk/nature/adaptations/Precocial`
    `http://en.wikipedia.org/wiki/Superprecocial`
  - Altricial species are helpless and need days, weeks, months to grow their software architectures

    (e.g. eagles, chimps, humans...)
    `http://www.bbc.co.uk/nature/adaptations/Altricial`
    `http://www.stanford.edu/group/stanfordbirds/text/essays/Precocial_and_Altricial.html`
  - These distinctions are criticised and modified in

    (Sloman & Chappell, 2005; Chappell & Sloman, 2007)

Why are precocial and altricial species so different?

Why are some born/hatched so incompetent (humans, apes, crows, hunting mammals), others able partly to fend for themselves (chicks, most invertebrates, horses, deer)?

- Abilities of newly hatched chickens and newborn deer, show that evolution can pre-design animals hatched or born more sophisticated than current robots.
  The young wildebeest does not have time to learn to run with the herd, if attacked by a predator.
- Compared with the task of walking and running on a grassy plain, some hunters, treetop-dwellers and berry pickers need an intricate grasp of spatial structure and motion (including skills needed to tear open a carcass to get at meat): but not all need the same understanding of their environment.
  E.g. rooks vs squirrels vs caterpillars in trees.
- If evolution cannot pre-design all the intricate mechanisms, it can, instead, use a bootstrapping architecture – as in "altricial" species.
- Compare the design requirements (niches) for adults of different species, including the need for some to be able to fit different niches (changing sets of requirements).

Perhaps we need different sets of concepts to describe what an adult lion sees and what an adult deer sees – if the affordances they can perceive and the actions they can perform are very different.

See `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#gibson`
What's vision for, and how does it work?
From Marr (and earlier) to Gibson and Beyond

Altricial species and cultures

Some individuals in altricial species develop by interacting with culturally determined environments (e.g. humans).

This provides scope for even more architectural variation in the resulting bootstrapped virtual machines:

- Different collections of perceptual hierarchies
- Different collections of thinking skills and formalisms (e.g. different human languages).
- Different collections of value systems
- Different decision-making architectures

Don't ask "what it is like" to be a human being born and bred in a totally different culture.

That's another variety of "anthropomorphism"!

Even within a culture, a mathematician's mind could have a partly different architecture from a dancer's, making some of their experiences when looking at the same things different in ways that neither can understand.

Architectural diversity

Within each architecture expect to find families of concepts where you previously thought there was one.

- different kinds of learning — MANY kinds
- many notions of consciousness (and qualia)
- different sorts of beliefs, intentions, desires
- different types of languages, different types of semantics

  Including ‘generalised languages’ (GLs) used primarily for thinking perceiving, formulating motives, planning, selecting goals, ... (don’t assume languages are needed only for communication).
- different sorts of emotions

  primary, secondary, tertiary emotions (and other types??)
- different kinds of moods, motivations, attitudes

COMPARE THE ARCHITECTURE OF MATTER

- the periodic table of the elements
- the variety of types of chemical compounds
- the variety of types of chemical processes

There is only one physical (chemical) world, whereas there are many types of minds, each supporting different collections of concepts of mentality.

What Kind Of Machine Can Have Emotions?

PROBLEM:

Many different definitions of “emotion”. in psychology, philosophy, neuroscience . . .
and many variants within each discipline

I’ve seen over 10 different definitions in research literature, and some researchers have found more.

DIAGNOSIS:

Different theorists concentrate on different phenomena.

We need a theory that encompasses all of them.

REPHRASE:

What are the architectural requirements for human-like mental states and processes?

Machines which have such architectures will be able to have human-like emotions.
(Unlike new born babies!)

Analysis of architectures with different information processing layers that evolved at different stages of evolutionary history can points to different classes of emotions linked to different layers in the architecture:

primary emotions – disruptive states in reactive control mechanisms,

secondary emotions – involving deliberative abilities to think about possible futures, possible causes

tertiary emotions – using meta-semantic and meta-management abilities to represent, reason about and manipulate processes and mechanisms in individuals with all these abilities, with many subdivisions.

There’s lots more work to be done analysing and comparing different sub-types.

How AI has changed

Early AI research was mainly about representations and algorithms.
In the mid-late 1980s attention switched to questions about architectures.

Compare (Sloman, 1978, Chapter 6) (Pylyshyn, 1991) (McCarthy et al., 2002) (Minsky, 2006), and many more.

Architectures are important because complex systems are made of concurrently active interacting sub-systems.

We need to learn about ways in which things can be put together.

But the space of architectures is enormous.

We can, however, see it as including various kinds of sub-architectures, including combinations of increasingly sophisticated functionality (e.g. three layers):

- REACTIVE
- DELIBERATIVE
- REFLECTIVE (SELF-MONITORING, SELF-CONTROLLING) ...

We can also divide the functionality in terms of relations to environments (three towers):

- SENSORY/PERCEPTUAL SYSTEMS
- INTERNAL PROCESSING
- MOTOR SYSTEMS

Superimposing both ways of subdividing parts of architectures produces the (very crude, but useful) CogAff generative schema, mentioned later.

That is not an architecture, but a framework for describing and comparing architectures.

(A sort of “grammar” for architectures.)

We need good organising ideas.

Many people produce architecture diagrams, and then tell stories about how they work,

but we need to look for good organising principles,

and we need to identify constraints to narrow the variety.
Obvious constraints:

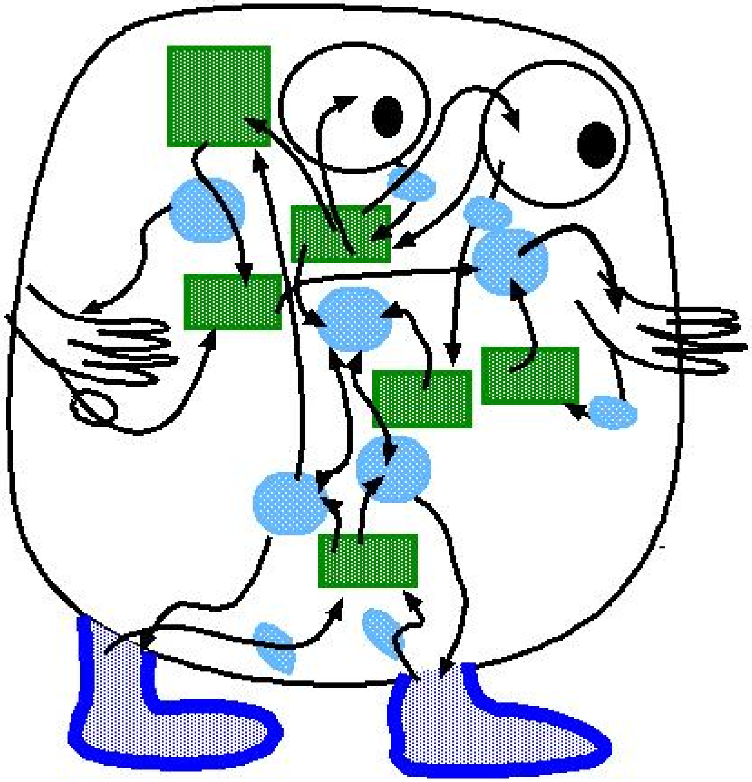
- physical possibility
- tractability
- being suited to required functionality
- being implementable in biological mechanisms
  (don't assume we know what they are!)

(Beware of *fashionable* constraints: groundedness, embodiment, situatedness ...)

More subtle constraints: "what is evolvable in various environments".

See also: (McCarthy, 2008)

Can Biological Evolution Produce An Unintelligible Mess?



Yes, in principle, but there is a counter argument...

Evolution and Engineers

It can be argued that evolution has similar requirements to engineers:

- Re-usable components

  "duplicate then differentiate" is common in evolution;

- Near decomposability

  so that a change in one place will not disrupt everything else

- Robust and general mechanisms
- Able to engage with our physical environment

But the requirements are different in different regions of design space and niche space:

The same physical environment can implicitly specify different requirements for improvement in different organisms living in that environment.

The Birmingham CogAff architecture schema provides a way of thinking about a wide variety of evolvable architectures.

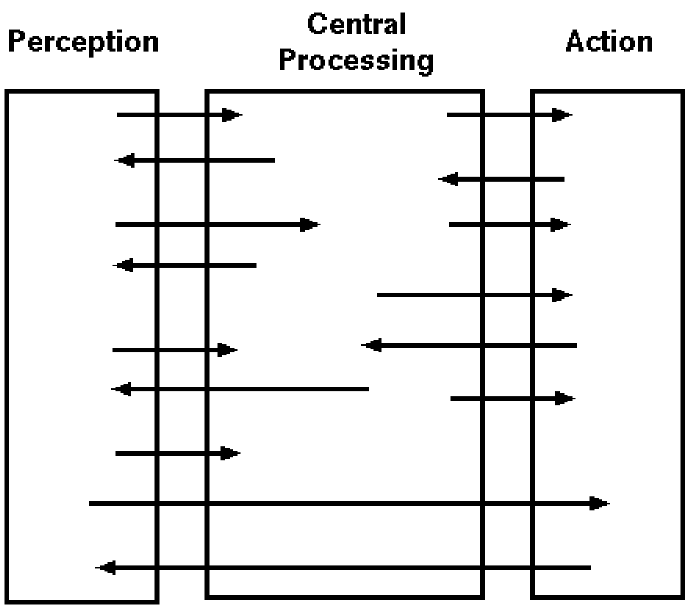
The CogAff schema came from thinking about architectures in two dimensions:

- Crude divisions between perceptual functions, central processing, and action (the three "towers" in (Nilsson, 1998)), and
- Crude divisions between types of sophistication produced at different evolutionary stages (the three "layers" in (Nilsson, 1998)).

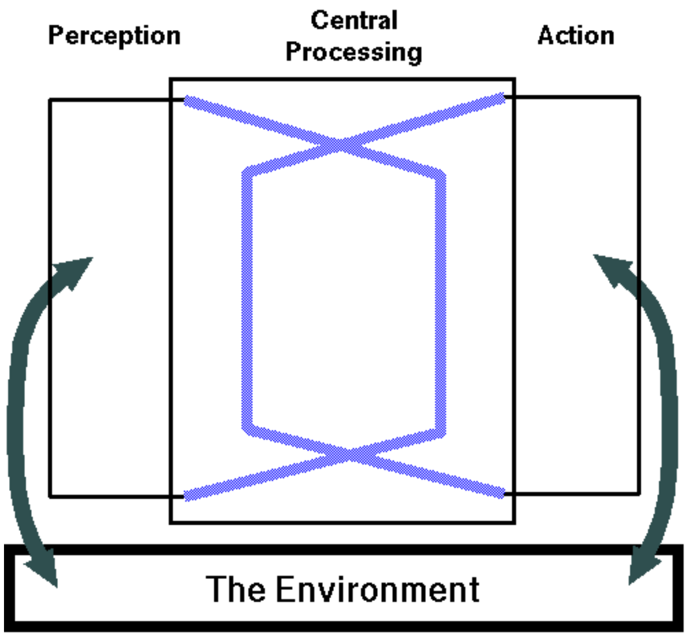
Superimposing the two dimensions gives a (messy) grid: different architectures use different parts.

Later we introduce H-Cogaff, a special sub-class covering human-like architectures using all of the grid.

The three towers – two versions



(a)



(b)

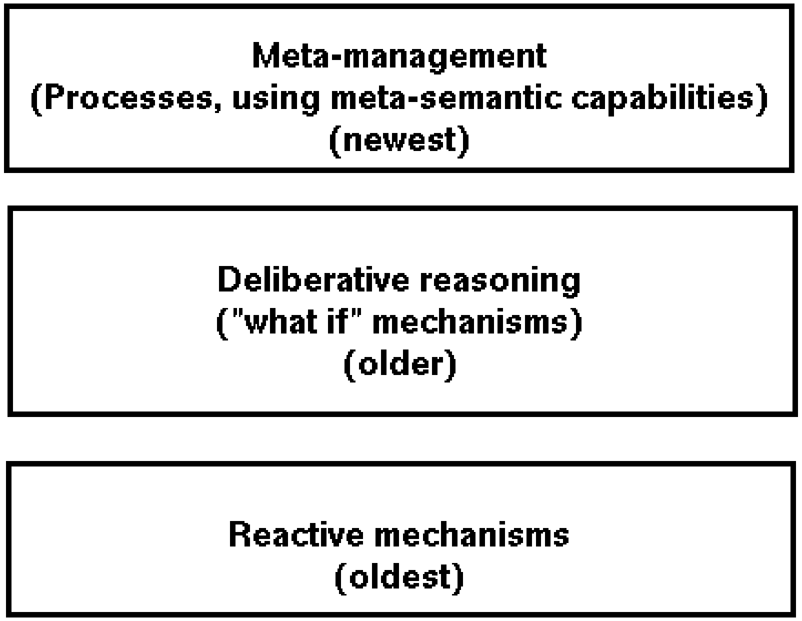
Version (a) assumes three fairly separate collections of sub-mechanisms, two of them connected with the environment, and each connected to a central collection of sub-mechanisms.

Version (b) assumes that the “towers” overlap and that both perception and action mechanisms can include two-way interaction with the environment.

E.g. perceptual processes such as tactile sensing and visual scanning integrate acting with perceiving, and many actions make use of feedback and other mechanisms in achieving fine-grained control.

For more on this see (Gibson, 1966).

The “three layer” View



Meta-management
(Processes, using meta-semantic capabilities)
(newest)

Deliberative reasoning
("what if" mechanisms)
(older)

Reactive mechanisms
(oldest)

On the three layered view the oldest organisms and the simplest artefacts use purely reactive architectures (bottom of diagram),

containing mechanisms that respond immediately (online, in real time) to whatever signals they are receiving, from the environment or from other parts of the system.

Later, biological evolution produced deliberative mechanisms

representing and reasoning about hypothetical situations and also past events, remote, invisible, portions of the environment, and possible actions (needs new forms of representation and new mechanisms).

Later still, organisms evolved with meta-semantic capabilities, required for representing and reasoning about things that themselves refer, represent, or reason, including oneself and other individuals in the environment – using mechanisms that support referential opacity.

The meta-management mechanisms are (to some extent) able to represent, reason about, manipulate and control the individual’s own internal states and processes concerned with information processing, including perception, motive generation, motive selection, planning, intention formation, and many more.

Different kinds of learning and development are associated with these different architectural layers.

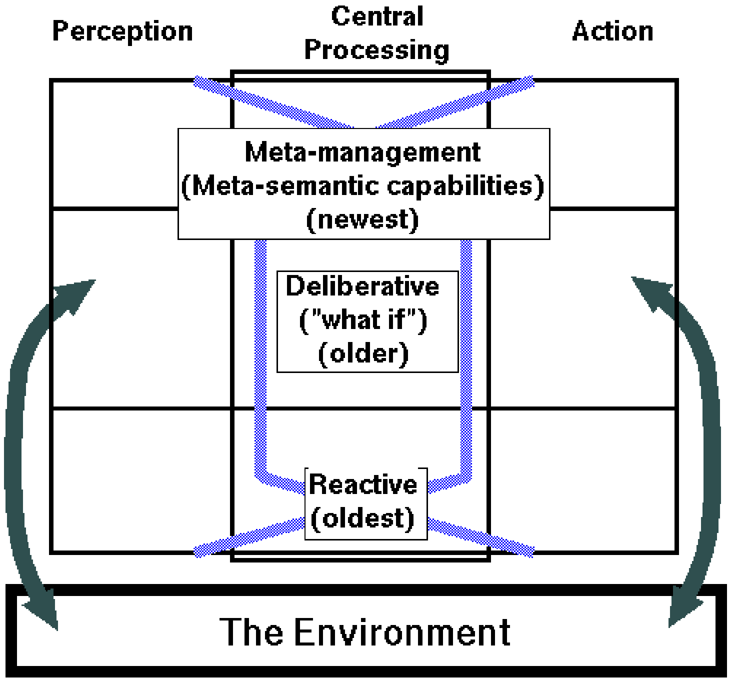
See also (Sloman, 2006). The “subsumption architecture” of (Brooks, 1986) uses several layers, but they are all purely reactive.

LAYERS + TOWERS = GRID

If we combine the layers idea and the towers idea we get the notion of a grid of of co-evolved concurrently active sub-mechanisms (almost sub-organisms) each contributing to the “niches” of the others.

Using the overlapping towers idea we get the second, more complex grid:

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



Perception

Central Processing

Action

Meta-management
(Meta-semantic capabilities)
(newest)

Deliberative
("what if")
(older)

Reactive
(oldest)

The Environment

The second grid is more realistic, as indicated above.

Different designs (different evolved organisms) will make use of different subsets of the grid, and may allow different connections between subsets.

Some may allow the various subsets to be grown or developed as a result of learning processes instead of being rigidly prescribed by the genome. (Sloman & Chappell, 2005) (Chappell & Sloman, 2007)

Notes on architectural layers

Different information-processing layers need not map onto different portions of the brain in any simple way

Mechanisms in a layer may be implemented in virtual machinery that does not inhabit a specific location.

Compare and contrast: MacLean's theory of the "triune" brain, built from these three layers:

reptilian, old mammalian, new mammalian

Also note the partially similar layered theory of James Albus in (Albus, 1981)

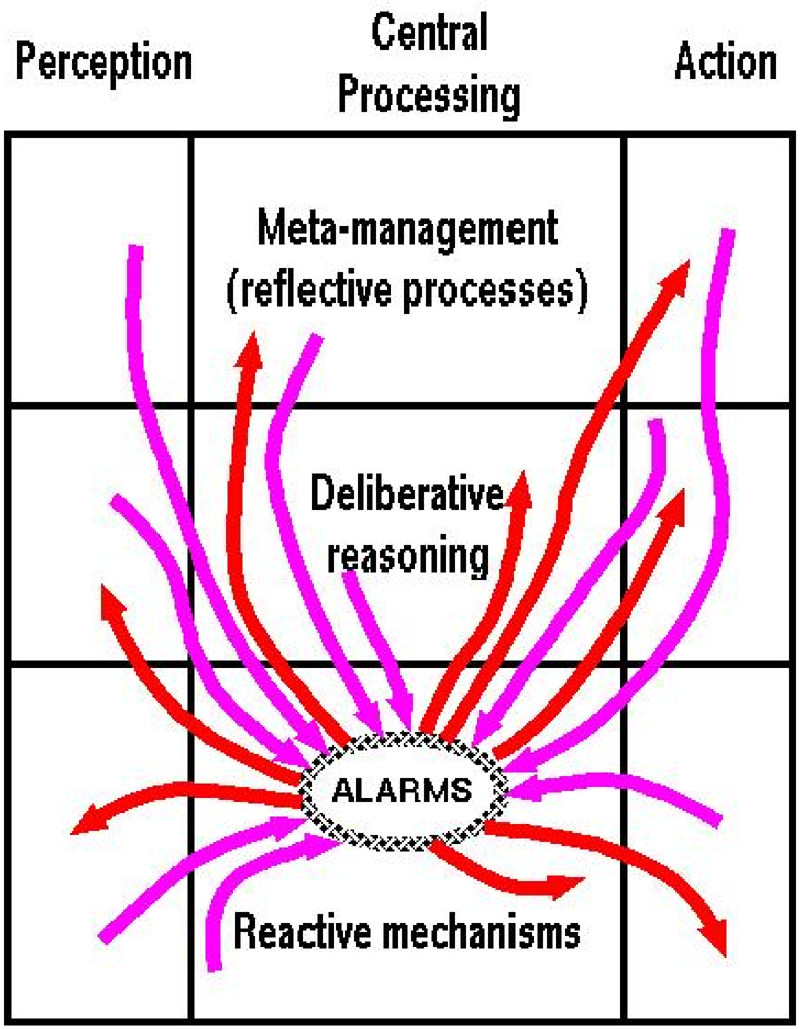
In contrast with those and other ideas, our layers do not form a dominance hierarchy.

Sometimes reactive mechanisms can control all of the others.

E.g. an "alarm mechanism" that detects dangers and opportunities requiring urgent action, may override all others.

One instance of the ‘CogAff’ architecture schema

A reactive “alarm” mechanism that can be triggered by inputs from any part of the system and can redirect or abort what is going on in any part of the whole system.



This is motivated by superimposing the ‘three tower’ (input-central-output) and ‘three layer’ (three stages of evolution) views depicted above – plus alarms, which are part of the reactive sub-system.

Note that this is NOT a dominance hierarchy. Control can go up as well as down: the systems work in parallel, influencing one another. Missing components are described elsewhere.

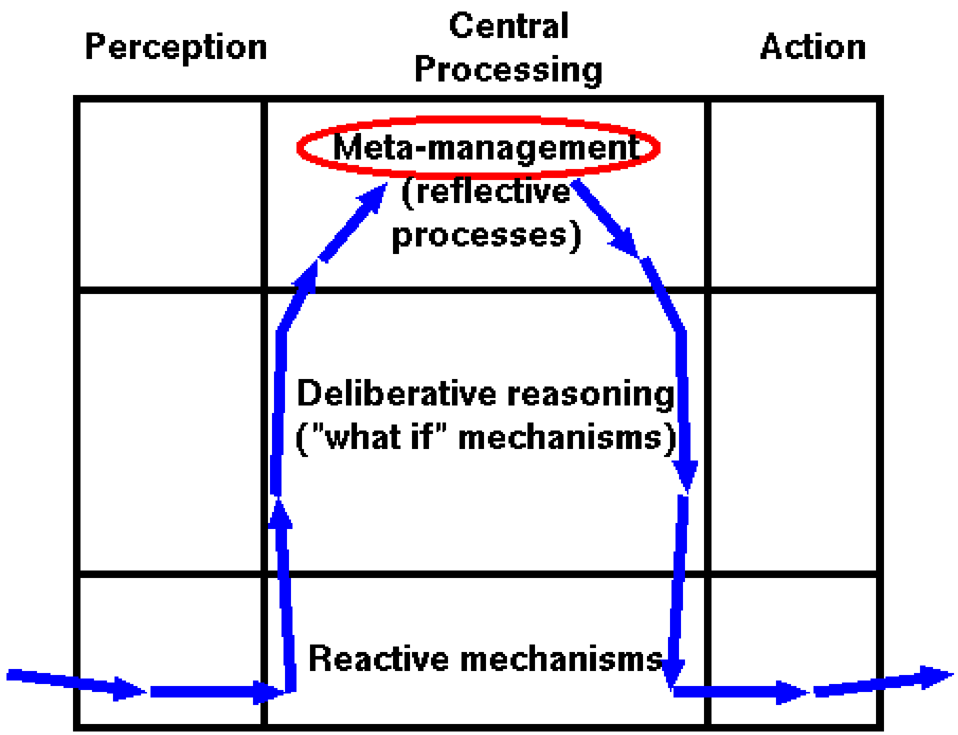
Layered architectures have many variants

With different subdivisions and interpretations of subdivisions, and different patterns of control and information flow.

Different principles of subdivision in layered architectures

- evolutionary stages
- levels of abstraction,
- control-hierarchy,
  (Top-down vs multi-directional control)
- information flow
  (e.g. the popular ‘Omega’ Ω model of information flow)

The “Omega” model of information flow


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

Rejects layered concurrent perceptual and action towers separate from central tower.

There are many variants, e.g. the “contention scheduling” model. (Shallice, Norman, Cooper)

Some authors propose a “will” at the top of the omega.

Shallice has elaborated the SAS: Supervisory Attention System at the ‘top’. The ideas overlap with meta-management.

CogAff allows more ‘horizontal’ connections.

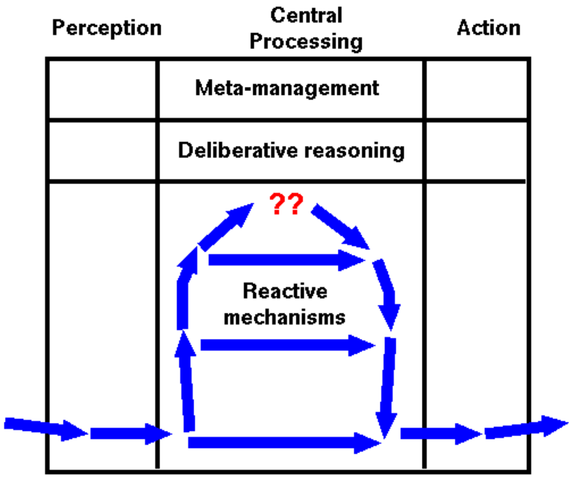
Most systems differ from the CogAff framework by not allowing the perception and action systems to include hierarchies of abstraction with direct connections at all levels to central layers. Hence the horizontal connections are only at the lowest level: ‘peephole’ vs ‘multi-window’ perception and action.

The contents of the different abstraction levels are discussed in other talks and papers here:

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

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

Another variant: Subsumption architectures(Brooks)



This allows layers of control, within the reactive category, but Brooks (sometimes) denies that animals use deliberative mechanisms.

His view appears to have changed recently (2002): http://204.194.72.101/www/oy8guwod/structure.pdf

CogAff is a schema: NOT ALL COMPONENTS ARE PRESENT IN ALL ANIMALS

(or all robots, all software agents)

What sort of architecture suffices for an insect?

Will a purely reactive architecture suffice?

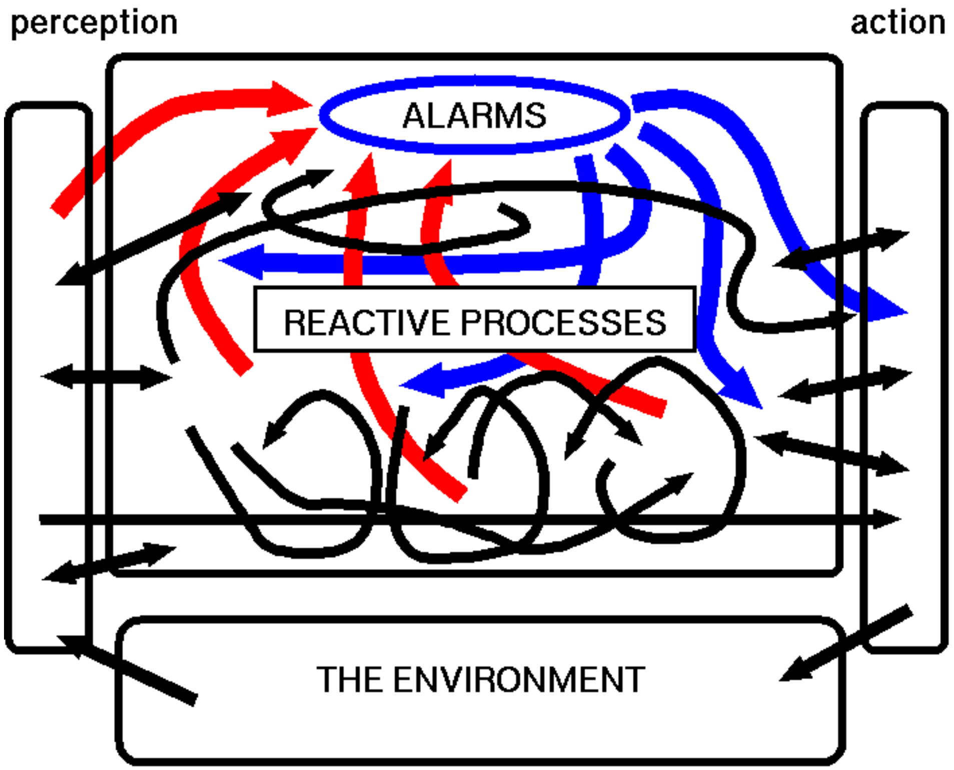
Can any insects do deliberation? Any fish? Any reptiles?

How many animals have a deliberative layer? E.g. mice, cats, eagles, monkeys, chimps?

Add meta-management for human-like systems. Chimps?

We can study the tradeoffs by exploring neighbourhoods in design space: what difference does it make if component X is added, or removed, or varied in some way?

EMOTIVE INSECTS?

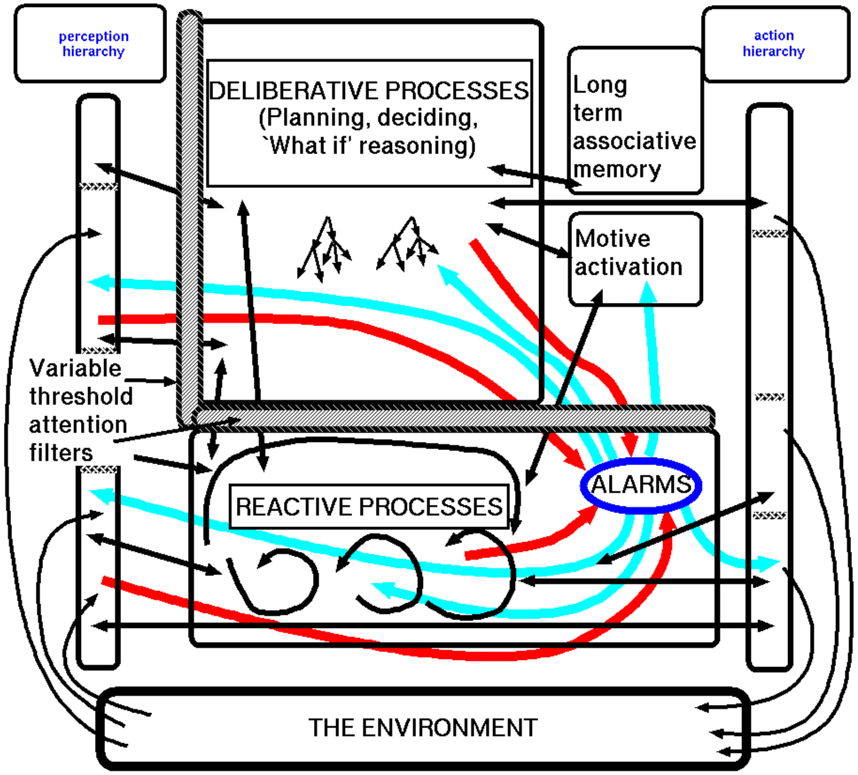


ALARM MECHANISM **(Global interrupt/override):**

- **Allows rapid redirection of the whole system, for sudden dangers or sudden opportunities**
- **FREEZING**
- **FIGHTING, ATTACKING**
- **FEEDING (POUNCING)**
- **GENERAL AROUSAL AND ALERTNESS (ATTENDING, VIGILANCE)**
- **FLEEING**
- **MATING**
- **MORE SPECIFIC TRAINED AND INNATE AUTOMATIC RESPONSES**

Related to what Damasio and Picard call: “Primary Emotions”

REACTIVE AND DELIBERATIVE LAYERS WITH ALARMS



How many animals combine reactive abilities with deliberative abilities, e.g. the ability to contemplate, evaluate, compare and choose between possible predictions regarding the actions of another, or possible plans for achieving some goal?

What are the architectural requirements for such capabilities?

Many requirements for hybrid systems still to be investigated

- What sort of long term memory (memories) SUPPORTING DIFFERENT KINDS OF DELIBERATION
- Different sources of motivation (EXTERNAL, INTERNAL, TRIGGERED BY BODILY NEEDS *vs* TRIGGERED BY THOUGHTS OF WHAT MIGHT HAPPEN)
- Attention filters for situations where motives are generated too fast to be processed properly
- Training of reactive layer by deliberative layer (PRODUCING CHANGES INDIRECTLY OVER A PERIOD OF TIME)

AN ALARM MECHANISM

(BRAIN STEM, LIMBIC SYSTEM?)

ALLOWS RAPID REDIRECTION

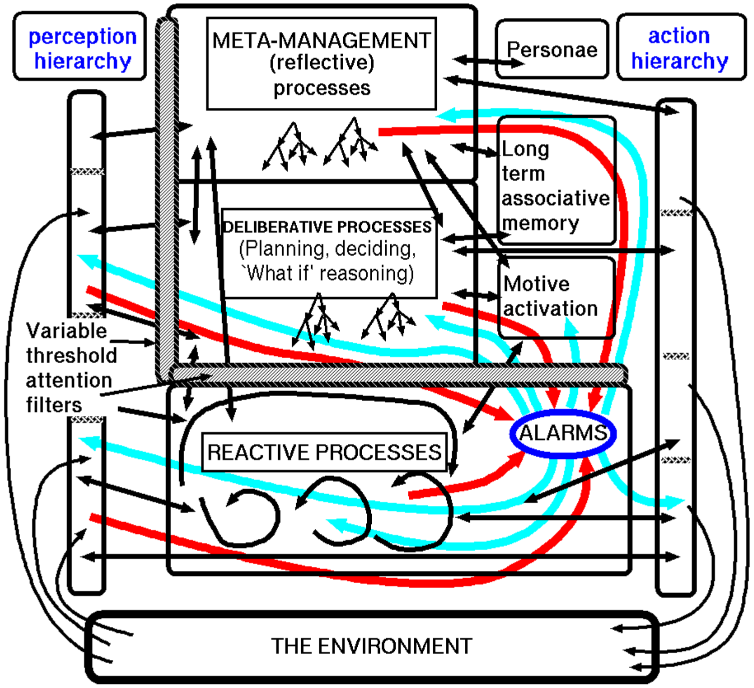
OF THE WHOLE SYSTEM.

But can be triggered by and can redirect deliberative processes.

ALARMS IN A HYBRID ARCHITECTURE

- **Freezing, fleeing, arousal etc. as before**
- **Becoming apprehensive about anticipated danger**
- **Rapid redirection of deliberative processes.**
- **Relief at knowing danger has passed**
- **Specialised learnt responses: switching modes of thinking.**

The H-Cogaff Architecture


perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, `What if' reasoning)
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

Human-like systems include meta-management and other evolutionarily recent additions.

A meta-management layer or reflective layer

This includes the ability to

- monitor,
- categorise,
- evaluate,
- (to some extent) redirect and modulate other internal processes.

But the third layer never has total control. Other parts of the system are concurrently active and potentially able to disrupt it.

Why? Because the environment is partly unpredictable

It can be disrupted by alarms, salient percepts, etc.

THE THIRD LAYER enables SELF-MONITORING, SELF-EVALUATION SELF-CONTROL (and qualia!)

This makes possible “tertiary” emotions, through having and losing control (of thoughts and attention:)

- Feeling overwhelmed with shame
- Feeling humiliated
- Aspects of grief, anger, excited anticipation, pride,
- Being infatuated, besotted and many more
  typically HUMAN emotions. (Contrast attitudes.)

Animals, infants, robots without a meta-management will not be able to have the typical human adult emotions described by poets, playwrights, gossips. But they may have other, older types.

Compare effects of different sorts of brain damage.

NOTES:

1. Different aspects of love, hate, jealousy, pride, ambition, embarrassment, grief, infatuation can be found in all three categories: primary, secondary and tertiary emotions.

2. Remember that these are not STATIC states but DEVELOPING processes, with very varied aetiology.

3. And they need yet more INTERNAL LANGUAGES

Beware of simplifications

(Systems may be “nearly decomposable”, and boundaries can change with learning and development).

Many variants: (NILSSON, ALBUS)

E.g. the towers may be thick or thin. They may have internal processing layers.

Both perception and action can be hierarchical, with multi-directional information flow.

WARNING:
Evolution, like other designers, can produce bugs

Some are hardware bugs, e.g. physical components with design infelicities (you can't sit in one position for a long time).

Some are control bugs, e.g. auto-immune diseases.

Some are software bugs, e.g.

- various kinds of psychiatric disorder,
- types of self-delusion,
- limitations of short-term memory or processing accuracy,
- buggy interrupt systems,
- many kinds of fallacious reasoning
- religious beliefs,
- nationalism,
- racism,
- overconfidence in one's own theories

Except in special circumstances, it is impossible to eliminate bugs in complex systems that interact with complex and changing, partly unknown, environments.

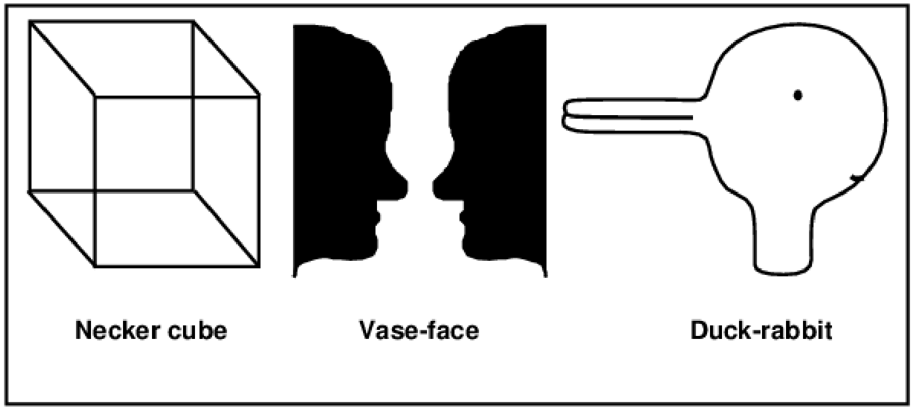
A good architectural theory should help to explain why various sorts of bugs are likely, and should help us search for them.

Could this be illustrated in research on psychotherapy, learning difficulties, addictions, and various kinds of social misfits?

Levels in perceptual mechanisms

Detecting: low level physical changes at transducers, remote entities, different varieties of segmentation, different levels of interpretation.

Seeing the switching Necker cube requires geometrical percepts.



Seeing the flipping duck-rabbit uses far more subtle and abstract percepts, going beyond geometric and physical properties. (Compare Marr on vision)

Things we can see besides geometrical properties:

- Which parts are ears, eyes, mouth, bill, etc.
- Which way something is facing
- Whether someone is happy, sad, angry, etc.
- Whether a painting is in the style of Picasso...
- Two faces holding a vase wedged between them!

See also
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/#talk7
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/#talk9

Extending Gibson's theory:

Evolution of perceptual mechanisms

Different perceptual sub-systems use different affordances, and different ontologies.
LIKE DIFFERENT ORGANISMS

Different levels of perceptual abstraction required for different purposes.

WHY?

To meet the more sophisticated requirements of more sophisticated co-evolved central components.

These in turn can evolve to make new uses of more sophisticated perceptual layers.

Likewise layered action systems.

A mind (or brain) is a co-evolved ecosystem.

See also:
A.Sloman (1989)
“On designing a visual system
(Towards a Gibsonian computational model of vision)”,
In *Journal of Experimental and Theoretical AI*, 289–337.

Multiple sources of control, with changing dominance relationships

If the different components are concurrently active, then they can be both receiving and transmitting information at all times, and information can go in many directions through many pathways in parallel.

Then no one layer dominates the rest (as in subsumption)

Reflexes and alarms are examples of control by lower level reactive mechanisms.

Training and development can add new arrows (new information links) as well as new components within the nine boxes.

Diagonal arrows e.g. from a high level perceptual layer to a low level reactive mechanism may be the result of training to achieve speed and fluency.

More sophisticated processing may be slower

Dangerously slow: so fast, powerful, “alarm systems” needed

Alarm systems will inevitably be pattern-based and stupid!

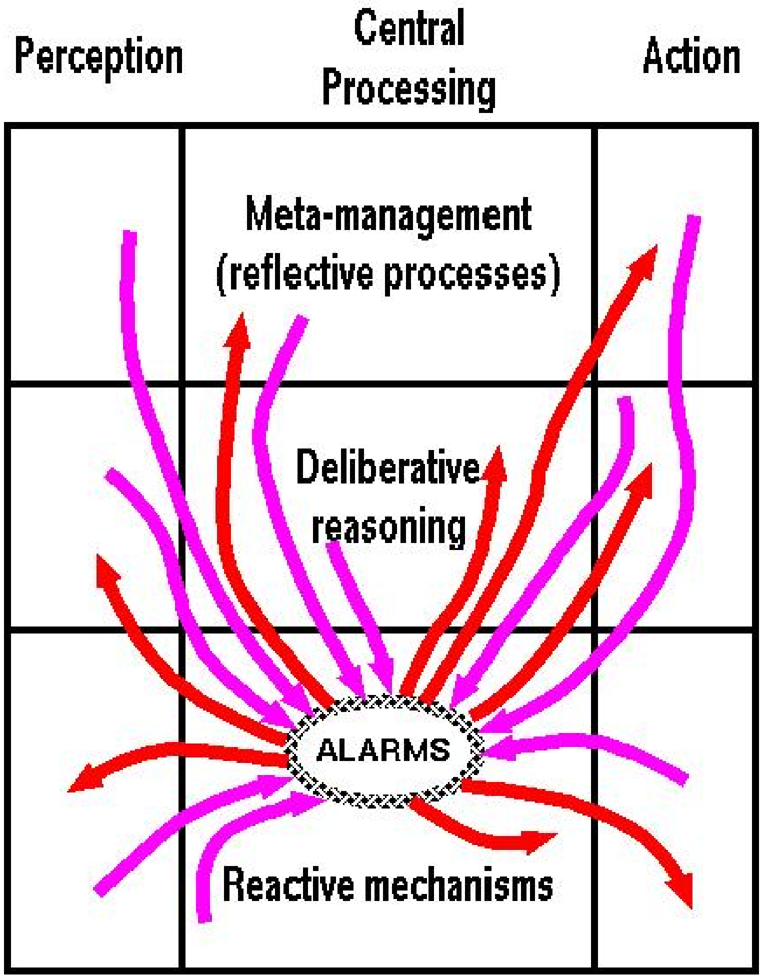
But they may be trainable.

There may be:

general global alarm systems, more local alarm systems and very specialised alarm systems (e.g. protective blinking reflex).

Alarm mechanisms are in the reactive layer (because of need for speed, which may also produce error!).

But they can have inputs from anywhere, and their effects can go anywhere (like the noise of fire-alarm bells in a building, even when it’s a false alarm!)



ADDITIONAL COMPONENTS

MANY PROFOUND IMPLICATIONS

e.g. for kinds of development

kinds of perceptual processes

kinds of brain damage

kinds of emotions

and other affective states

EXTRA MECHANISMS NEEDED

personae (variable personalities)

attitudes

standards & values

formalisms

categories

descriptions

moods (global processing states)

motives

motive comparators

motive generators (Frijda's "concerns")

Long term associative memories

attention filter

skill-compiler

The need for “inner languages”

All the different sorts of mechanisms need or process information.

They all need vehicles for the information.

They all therefore use “languages” of some sort.

Clearly in this sense internal languages for perceiving, learning, deliberating, thinking, desiring, etc. evolved long before external languages of the sort we now refer to as “languages”.

A more detailed analysis would take us into dimensions of variation of types of language (or representation), their syntax, their semantics, their pragmatics.

For more on evolution of language, see:

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`

Primary and secondary emotions in a hybrid architecture

Damasio & Picard:

Cognitive processes trigger “secondary emotions”.

From an architectural standpoint we can distinguish several different sub-categories of emotions:

E.g. *purely central* and *partly peripheral* secondary emotions.

On some (misguided) theories, the former are impossible!

When we add the meta-management layer, we find scope for another class “tertiary emotions”.

Thinking about too narrow a range of architectures (or not thinking about architectures) can hamper the search for explanatory theories.

There are many papers on all this in the Cogaff directory:

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

WE CAN EXPLAIN SOME DISPUTES AND CONFLICTING DEFINITIONS

E.G. of “emotion” “learning” “executive function” etc.

Different researchers focus on different features of a very complex system.

But they are unaware of the other features.

Like the proverbial collection of blind men all trying to say what an elephant is:

– One feels the trunk

– One feels a tusk

– One feels an ear

– One feels a leg

– One feels the tail, etc.

Each is correct — about a tiny part of reality. But they all draw wrong conclusions about the whole thing.

Could computer-based robots have all this?

Maybe. We don’t know enough yet about what the requirements are, or what computers and and cannot do.

Beware of spurious arguments: e.g.

- they could still be “zombies”
  (not with all that virtual machine architecture at work)
- brains use chemistry, whereas computers don’t.
- brains change continuously, computers are digital
- computers do only what they are programmed to do
  (said by people who have never programmed computers)
- minds need to be based on metabolism
  (but that’s just a very fine grained concurrent architecture)
- Gödel’s incompleteness theorem
  (a long, long story of philosophical muddle and delusion, based on superb mathematics)
- Only quantum non-local processes can explain mentality
  (maybe: but where exactly are they required in the architecture?)

WE DO NOT YET UNDERSTAND MUCH ABOUT ARCHITECTURES

- how many types they are
- what the trade-offs are
- how they evolve and develop
- how they differ among animals
- how they can be combined
- how different sorts can coexist in hybrid systems
  and how many concurrent processing pathways
  result from that
- how many kinds of action control there are
  and how they interact
- how many kinds of learning there are
  (Architecture-based concepts of learning)

CONCLUSION: THE SCIENCE

- Much of this is conjectural – many details still have to be filled in and consequences developed (both of which can come partly from building working models, partly from multi-disciplinary empirical investigations).
- An architecture-based ontology can bring some order into the morass of studies of affect (e.g. myriad definitions of “emotion”).

  Compare the relation between the periodic table of elements and the architecture of matter.
- This can lead to a better approach to comparative psychology, developmental psychology (the architecture develops after birth), and effects of brain damage and disease.
- It will provide a conceptual framework for discussing which kinds of emotions can arise in software agents that lack the reactive mechanisms required for controlling a physical body.

CONCLUSION: ENGINEERING

Designers need to understand these issues:

(a) if they want to model human affective processes,

(b) if they wish to design systems which engage fruitfully with human affective processes,

(c) if they wish to produce teaching/training packages for would-be counsellors, psychotherapists, psychologists.

(d) and maybe even for convincing synthetic characters in computer entertainments?

FOR SCIENCE AND ENGINEERING:

Consider an ‘eco-system of mind’ rather than just a ‘society of mind’.

PHILOSOPHY OF MIND

WILL NEVER BE THE SAME AGAIN

Except that too many philosophers doing philosophy of mind have no idea how to design, test, debug and fix a working system.

Or how to specify requirements for a complex working system.

COGNITION and AFFECT PROJECT

PAPERS:

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

TOOLS:

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

(Including the SIM_AGENT toolkit)

`http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html`

THESE AND RELATED SLIDES CAN BE FOUND IN

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

THE END(for now)

Some slides not presented at the talk follow

SENSING AND ACTING CAN BE ARBITRARILY SOPHISTICATED

- Don't regard sensors and motors as mere transducers.
- They can have sophisticated information processing architectures.

  E.g. perception and action can be hierarchically organised with concurrent interacting sub-systems.

- Perception goes far beyond segmenting, recognising, describing what is "out there". It includes:
  - providing information about *affordances* (Gibson, not Marr, but co-evolved beasties better)
  - directly triggering physiological reactions (e.g. posture control, sexual responses)
  - evaluating what is detected,
  - triggering new motivations
  - triggering "alarm" mechanisms
  - . . . . .

And these all need languages of some sort

THE META-MANAGEMENT LAYER NEED NOT HAVE CONSTANT CONTENTS

Different ‘personalities’ (personae) in different contexts

- At home with the family
- Driving on a motorway
- Interacting with subordinates at work
- Being interviewed by superiors
- In the pub with chums

...and many more ...

WHERE CONTROL BY A PERSONALITY INVOLVES TURNING ON A LARGE COLLECTION OF:

- skills,
- styles of thought and action,
- types of evaluations,
- decision-making strategies,
- reactive dispositions,
- ....

COMPARE THE MUCH FASTER GLOBAL CHANGES PRODUCED BY ALARM MECHANISMS: PERHAPS AN EVOLUTIONARY PRE-CURSOR OF METAMANAGEMENT?.

The meta-management system is a framework which can be occupied by different 'control regimes' at different times?

THIS REQUIRES

- A store of 'personalities'
- Mechanism for acquiring and storing new ones
  and modifying extending old ones
- Mechanisms for 'switching control' between
  personalities.

WHAT FOR?:
Different contexts have different requirements.

Global switching triggered by context may be more effective than always having to select individual rules, strategies, information items etc. on the basis of

TASK + LOCAL CONTEXT + GLOBAL CONTEXT

In people, switching personality is often involuntary and even unconscious (i.e. unnoticed).

WHY?

Can we learn to be more self-aware? What needs to change?

META-MANAGEMENT AND SOCIAL CONTROL

A SOCIETY OR CULTURE CAN INFLUENCE INDIVIDUALS

E.G. by

- Training reactive mechanisms
  e.g. using reinforcement learning.
- Enabling successful plans, strategies, etc. to be
  transferred without having to be rediscovered.
- Training modes of coordination in collaborative
  activities,
- Transferring powerful formalisms
- Transferring useful modes of categorisation,
  ontologies (including ontologies of mental phenomena)
- Influencing evaluation mechanisms
  including evaluating internal events, actions
  (e.g. I was selfish, selfless, brave, stupid, wise, lucky)

This can be useful or harmful:
e.g. religious indoctrination which produces guilt about natural healthy desires, etc.

SOCIALLY IMPORTANT HUMAN EMOTIONS

INVOLVE RICH CONCEPTS
AND KNOWLEDGE

and

RICH CONTROL MECHANISMS
(architectures)

- Our everyday attributions of emotions, moods, attitudes, desires, and other affective states implicitly presuppose that people are information processors.
- To long for something you need to know of its existence, its remoteness, and the possibility of being together again.
- Besides these *semantic* information states, longing also involves *control* states.

  ONE WHO HAS DEEP LONGING FOR X DOES NOT MERELY OCCASIONALLY THINK IT WOULD BE WONDERFUL TO BE WITH X. IN DEEP LONGING THOUGHTS ARE OFTEN *uncontrollably* DRAWN TO X.
- Physiological processes (outside the brain) may or may not be involved. Their importance is normally over-stressed by experimental psychologists under the malign influence of the James-Lange theory of emotions. (Contrast Oatley, and poets.)

VARIETIES OF MOTIVATIONAL SUB-MECHANISMS

MOTIVATION IS NOT JUST ONE THING

Motives or goals can short term, long term, permanent.

They can be triggered by physiology, by percepts, by deliberative processes, by metamanagement.

So there are many sorts of motive generators: MG

However, motives may be in conflict, so motive comparators are needed: MC.

But over time new instances of both may be required, as individuals learn, and become more sophisticated:

Motive generator generators: MGG

Motive comparator generators: MCG

Motive generator comparators: MGC

and maybe more:

MGGG, MGGC, MCGG, MCGC, MGCG, MGCC, etc ?

THERE ARE ALSO “EVALUATORS”

The need for evaluators:

- Current state can be evaluated as good, or bad, to be preserved or terminated.
- These evaluations can occur at different levels in the system,
- and in different sub-systems,
- accounting for many different kinds of pleasures and pains.

  (OFTEN CONFUSED WITH EMOTIONS.)

Where are the motive generators and evaluators?

All over the system – not just at the ‘top’

(Contrast the Omega model of information flow.)

META-MANAGEMENT AND TERTIARY EMOTIONS

Tertiary emotions (previously called “perturbances”) involve interruption and diversion of thought processes.

I.e. the metamanagement layer does not have complete control.

WHY?

- New information from other sub-systems can cause interrupts.
- New motives from other subsystems can cause interrupts.
- Global alarm signals triggered by events elsewhere can cause interrupts and re-direction.

VARIABLE THRESHOLD INTERRUPT FILTERS CAN HELP REDUCE THESE EFFECTS.

Sometimes meta-management seems to be ‘turned off’, e.g when we are totally absorbed in some task.

QUESTION:
Is it essential that all sorts of emotions have physiological effects outside the brain, e.g. as suggested by William James?

NO: which do and which do not is an empirical question, and there may be considerable individual differences.

Some tertiary emotions may be purely central.

Different architectural layers support different sorts of mental phenomena and help us define
an architecture-based ontology of mind

Different animals will have different mental ontologies

Humans at different stages of development will have different mental ontologies

The REACTIVE layer with GLOBAL ALARMS supports “primary” emotions:

- being startled
- being disgusted by horrible sights and smells
- being terrified by large fast-approaching objects?
- sexual arousal? Aesthetic arousal ?

etc. etc.

The DELIBERATIVE layer enables “secondary” emotions (cognitively based):

- being anxious about possible futures
- being frustrated by failure
- excitement at anticipated success
- being relieved at avoiding danger
- being relieved or pleasantly surprised by success

etc. etc.

Things to read

References

Albus, J. (1981). *Brains, behaviour and robotics*. Peterborough, N.H.: Byte Books, McGraw Hill.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, *RA-2*, 14–23. (1)

Chappell, J., & Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. *International Journal of Unconventional Computing*, *3*(3), 211–239. (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609)

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Kant, I. (1781). *Critique of pure reason*. London: Macmillan. (Translated (1929) by Norman Kemp Smith)

McCarthy, J. (2008). The well-designed child. *Artificial Intelligence*, *172*(18), 2003-2014. Available from `http://www-formal.stanford.edu/jmc/child.html`

McCarthy, J., Minsky, M., Sloman, A., Gong, L., Lau, T., Morgenstern, L., et al. (2002). An architecture of diversity for commonsense reasoning. *IBM Systems Journal*, *41*(3), 530–539. (http://www.research.ibm.com/journal/sj41-3.html)

Minsky, M. L. (2006). *The Emotion Machine*. New York: Pantheon.

Nilsson, N. (1998). *Artificial intelligence: A new synthesis*. San Francisco: Morgan Kaufmann.

Pylyshyn, Z. (1991). The role of Cognitive Architectures in the Theory of Cognition. In K. VanLehn (Ed.), *Architectures for Intelligence* (pp. 189–223). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). Available from `http://www.cs.bham.ac.uk/research/cogaff/crp`

Sloman, A. (2002). Architecture-based conceptions of mind. In P. Gärdenfors, K. Kijania-Placek, & J. Woleński (Eds.), *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)* (pp. 403–427). Dordrecht: Kluwer. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#57`

Sloman, A. (2006, May). *Requirements for a Fully Deliberative Architecture (Or component of an architecture)* (Research Note No. COSY-DP-0604). Birmingham, UK: School of Computer Science, University of Birmingham. Available from `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`

Sloman, A., & Chappell, J. (2005). The Altricial-Precocial Spectrum for Robots. In *Proceedings IJCAI'05* (pp. 1187–1192). Edinburgh: IJCAI. (http://www.cs.bham.ac.uk/research/cogaff/05.html#200502)

And much more...