Oxford 22 Jan 2001

VARIETIES OF EVOLVABLE MINDS

How to think about architectures for human-like and other agents

OR

How to Turn Philosophers of Mind into Engineers

Aaron Sloman

School of Computer Science
The University of Birmingham

http://www.cs.bham.ac.uk/~axs/
A.Sloman@cs.bham.ac.uk

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

These slides are available online here
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/#talk1
Last modified 19 Jan 2003

IDEAS DEVELOPED IN COLLABORATION WITH

Steve Allen, Luc Beaudoin,
Darryl Davis, Catriona Kennedy,
Brian Logan, Matthias Scheutz,
Ian Wright,

and others in the

BIRMINGHAM COGNITION AND AFFECT PROJECT

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

I have, of course, also learnt from many others, e.g.

Margaret Boden, Marvin Minsky, Pat Hayes, Max Clowes,

to name a few.

OVERVIEW:
How to undo fragmentation

- Consider whole architectures
  (Not just language, vision, learning ...)
- Consider different species (not just humans)
- Consider individual differences (infants, brain-damaged...)
- Consider artefacts (not biological systems)
- Consider design requirements (and how they change)
- Consider design possibilities (beyond the obvious)
- Consider developmental and evolutionary trajectories
- Combine multiple disciplines (including philosophy)
  (Switch modes of thinking often)
- Acknowledge conceptual confusion
  (We don't necessarily mean what we think we mean)

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

“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 “epiphenomena”

They are “emergent” phenomena with causal powers.

But nothing spooky! Engineers design some of them.

Ask the Ghost of Gilbert Ryle

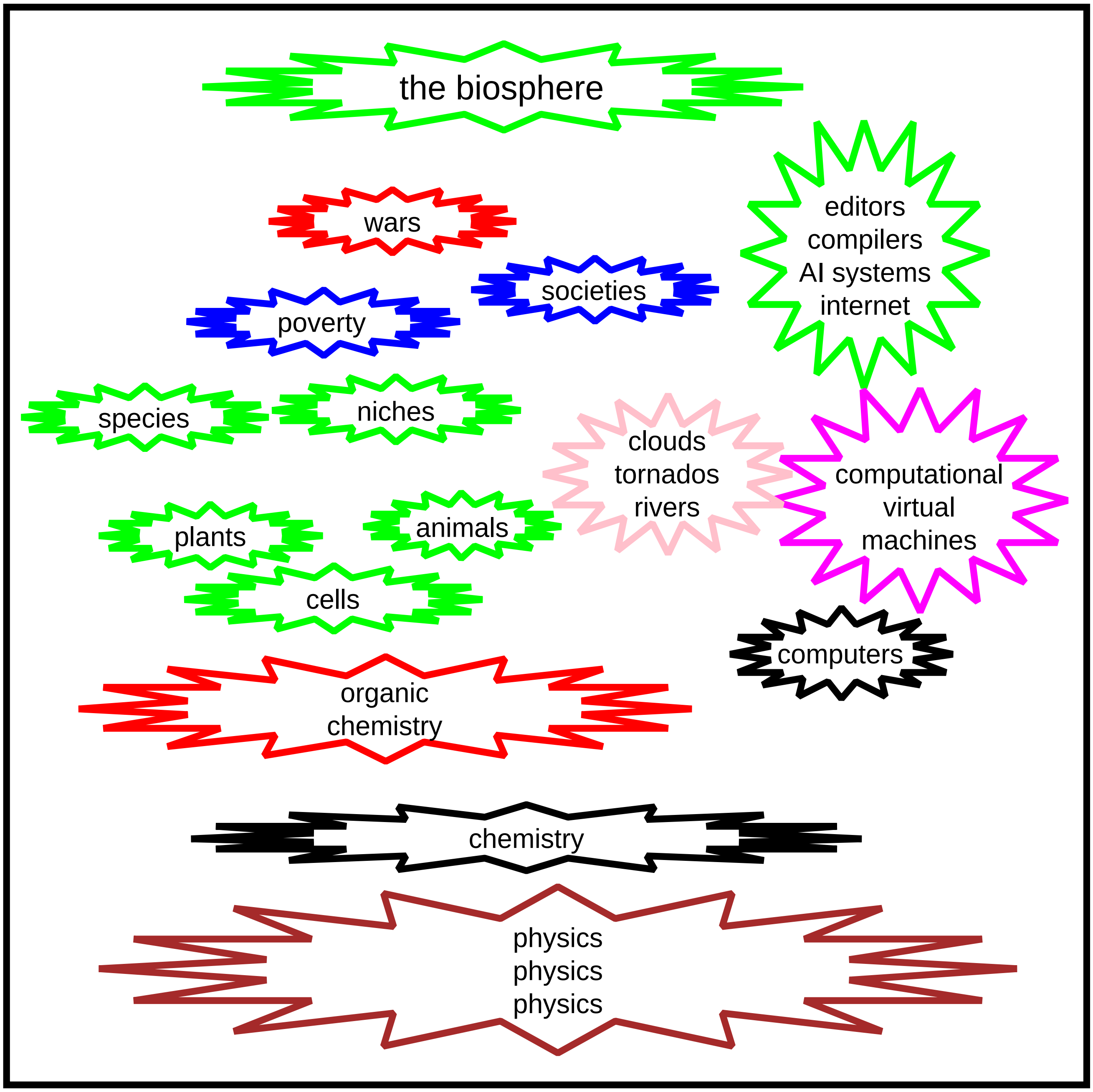
What sort of machine?

An information-processing machine.

Not necessarily a physical machine.

But it must be implemented in physical mechanisms if it is to DO anything

Emergent virtual machines are everywhere



How many levels of physics will there be in 500 years time?
Different levels involve different ontologies.

ENGINEERS AS PHILOSOPHERS

A common comparison:

$$\text{MIND} \Longleftrightarrow \text{BRAIN}$$

$$\text{VIRTUAL MACHINE} \Longleftrightarrow \text{PHYSICAL MACHINE}$$

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

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

The latter, however is very complex, and hard to make precise.

We need to distinguish different sorts of supervenience: property supervenience, pattern supervenience, agglomerative supervenience, and

Mechanism supervenience:
one ontology including causal interactions supervenes on another

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

Different VM architectures are required 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!

Let’s look at neighbourhoods and trade-offs

- **in design space**
- **in niche space**

Let’s analyse:

- **different types of *trajectories* through these spaces, in evolution, in individual development, in learning, in cultural change, in repairing, bug-fixing ...**
- **the interactions between the trajectories, i.e. *the many feedback loops* in co-evolution.**
- **architectures not only for individuals, but for sub-mechanisms and for larger structures:**

FAMILIES, TEAMS, PAIRS FIGHTING, ECONOMIC SYSTEMS, ECO-SYSTEMS.

No bit of this will be fully understood without putting it in the context of the rest.

IS EVOLUTION A DESIGNER?

Yes insofar as it produces designs:

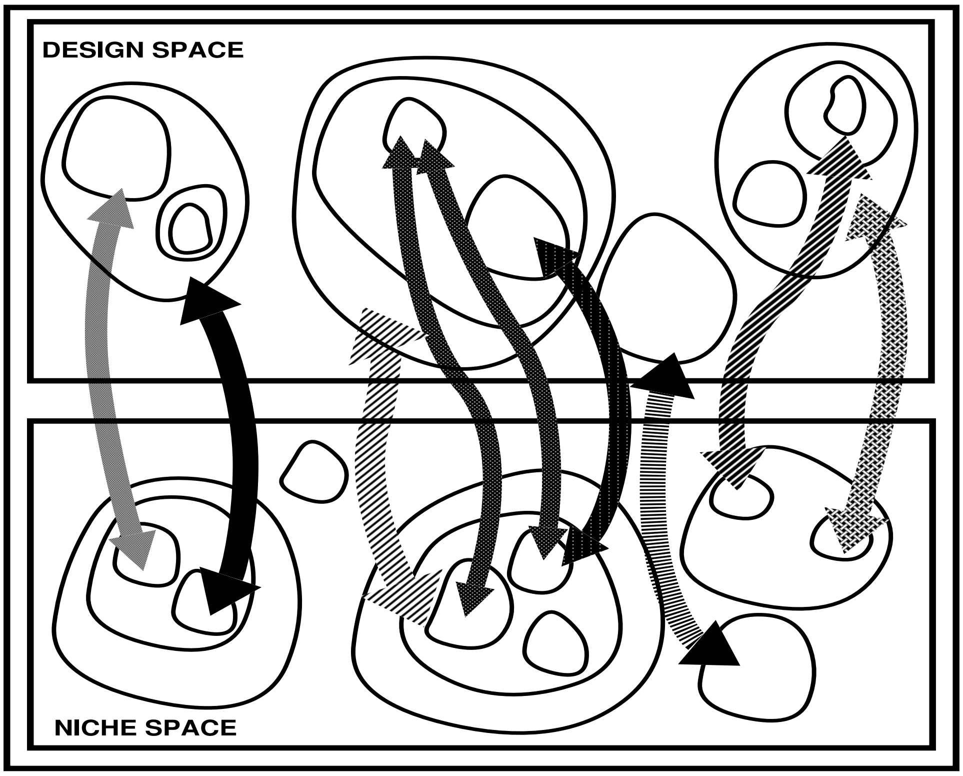
- Partly implicitly by producing instances of those designs
- Partly by producing re-usable specifications for designs in a powerful formalism (which we only partly understand)
- Also in using information in the process: information that is mostly scattered among all the co-existing, co-evolving species (information about varieties of environments, and what does and does not work in various environments.)

But it is a “reactive” system, not a “deliberative” system, in the sense defined later. It also lacks “meta-management”.

A possible exception: evolution can use the cognitive abilities of “intelligent” informed individuals, e.g. in mate selection.

Evolution produces *niches* as well as *designs*

DESIGN SPACE AND NICHE SPACE



Relations between designs and requirements (niches) are not just "fitness functions". They are multi-dimensional relationships. (Like *'Which?'* evaluations.)

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

There are different sorts of trajectories through the two spaces

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.

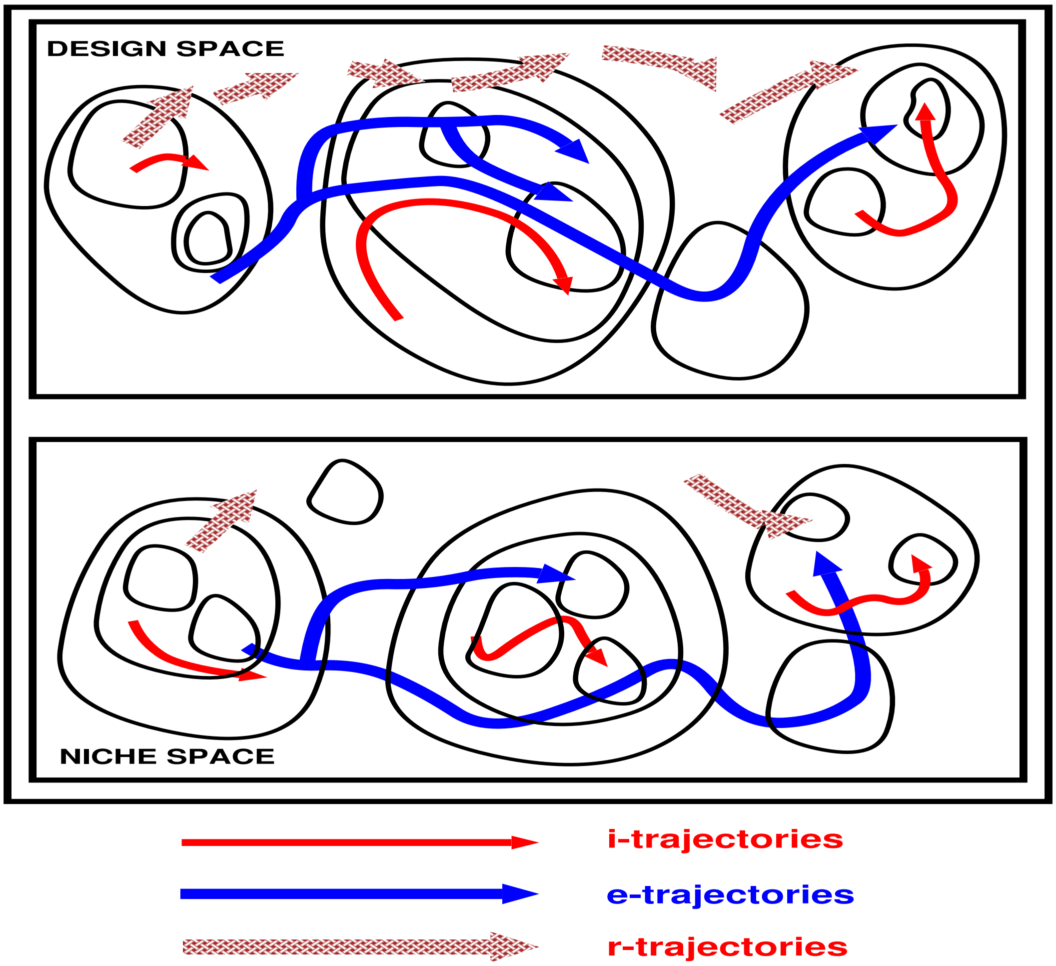
r-trajectory: possible for a system being repaired or built by an external designer whose actions turn non-functioning part-built systems into functioning wholes.

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.

All but r-trajectories are constrained by the requirement for "viable" systems at every stage.

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



An external "repairer" can push something through an "r-trajectory" (in which intermediate forms need not be viable.)

But don't forget that biological evolution is discontinuous.

Biological evolution: Multiple interacting e-trajectories, later using i-trajectories, then s-trajectories and c-trajectories, and now also r-trajectories (genetic engineering?)

Many questions: e.g. why are there so few “intelligent” species or individuals. (Count species, individuals or biomass.)

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

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.

EVOLUTION OF MIND
Different mental concepts are applicable in different architectures

An architecture supports a collection of possible states, processes, causal interactions:

Different collections for different architectures.

If mental concepts are architecture-based then we can't apply the same ones (e.g. ours) to all organisms.

Compare:

- **A fly that is "conscious" of my rapidly approaching hand**
- **An adult human "conscious" of a rapidly approaching mugger's fist**

Do not expect to be able to use your concepts to understand "What it is like" to be a fly, a bat a new born baby.

Perhaps evolution designed babies with the ability to fool parents into treating them as humans while they build their human architecture?

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.

Others cannot. How they perceive their environment will be importantly different.

- **Precocial species are born or hatched ready to feed, walk, swim, run, etc. (e.g. chickens, deer, horses...)**
- **Altricial species are helpless and need days, weeks, months to grow their software architectures (e.g. eagles, chimps, humans...)**

Why are precocial and altricial species so different?

Compare the design requirements (niches) for adults.

Compared with the task of walking and running on a grassy plain, hunters, treetop-dwellers and berry pickers need an intricate grasp of spatial structure and motion: but not all need the same grasp.

If evolution cannot pre-design all the intricate mechanisms, it can, instead, use a bootstrapping architecture.

So we need different sets of concepts to describe what a lion sees and what a deer sees.

Some individuals in altricial species develop by interacting with culturally determined environments

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
- 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 different architecture from a dancer's.

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
- different sorts of emotions

  primary, secondary, tertiary emotions (and more to come)
- 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

But 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

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!)

Our work points to at least three classes of emotions linked to different layers in the architecture which evolved at different times: *primary*, *secondary* and *tertiary* emotions, along with moods and other affective states.

AI used to be mainly about representations and algorithms
Now questions about architectures are equally (or more) important

We need to know how to put things together, but the space of architectures is enormous.

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

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

We can also divide the functionality (three towers):

- SENSORY/PERCEPTUAL SYSTEMS
- INTERNAL PROCESSING
- MOTOR SYSTEMS

We need some 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.**

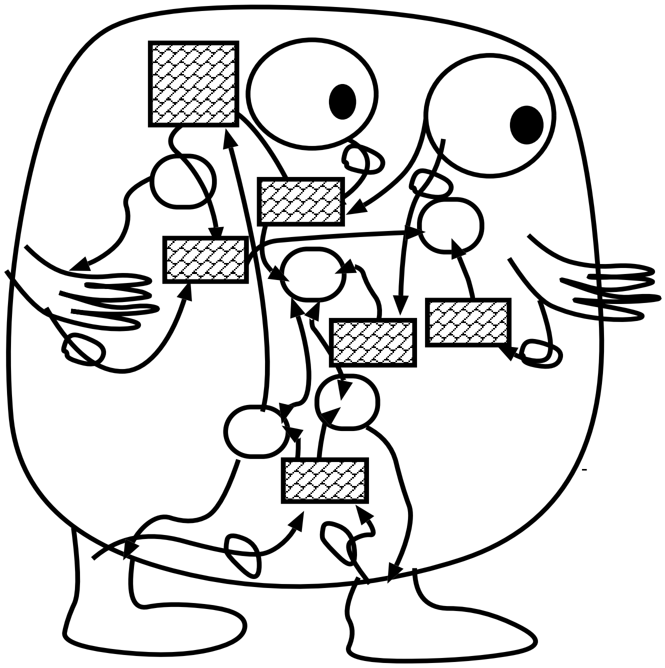
Obvious constraints:

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

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

More subtle constraint: "what is evolvable".

CAN BIOLOGICAL EVOLUTION PRODUCE AN UNINTELLIGIBLE MESS?



Yes, in principle.

However, it can be argued that evolution has similar requirements to engineers:

- **Re-usable components**
  ("duplicate then differentiate" is common)
- **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.

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

Later we introduce H-Cogaff, a special sub-class covering human-like architectures.

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).

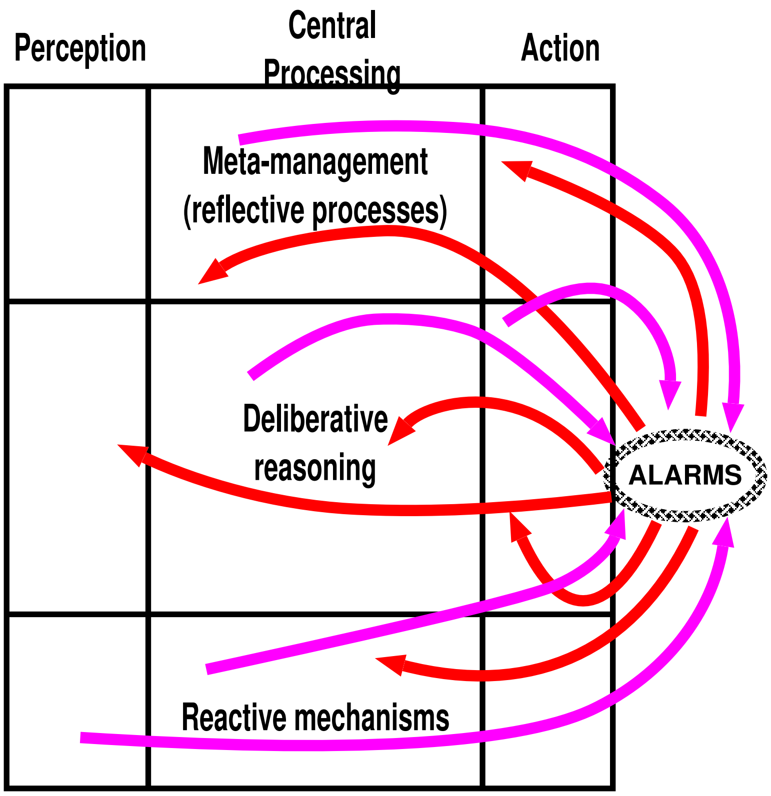
Some are control bugs, e.g. auto-immune diseases.

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**

It is impossible to eliminate bugs in complex systems. Our theories help to explain why some are likely.

The ‘CogAff’ Architecture Schema
(A partial view)

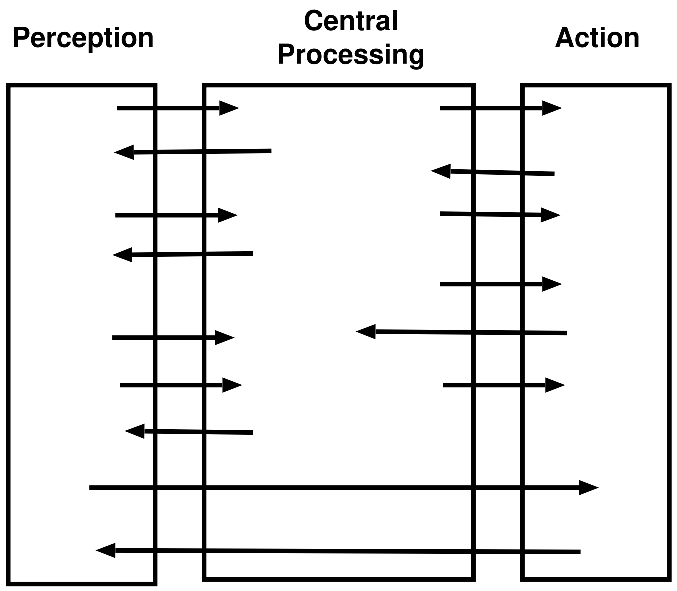


This is motivated by superimposing the ‘triple tower’ (input-central-output) and ‘triple layer’ (three stages of evolution) views depicted below – plus alarms, explained later (actually part of the reactive sub-system).

Note that this is NOT a dominance heirarchy. Control can go up as well as down: the systemd work in parallel, influencing one another.

Missing additional components are described later.

The “triple tower” View



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

Many variants: (NILSSON, ALBUS)

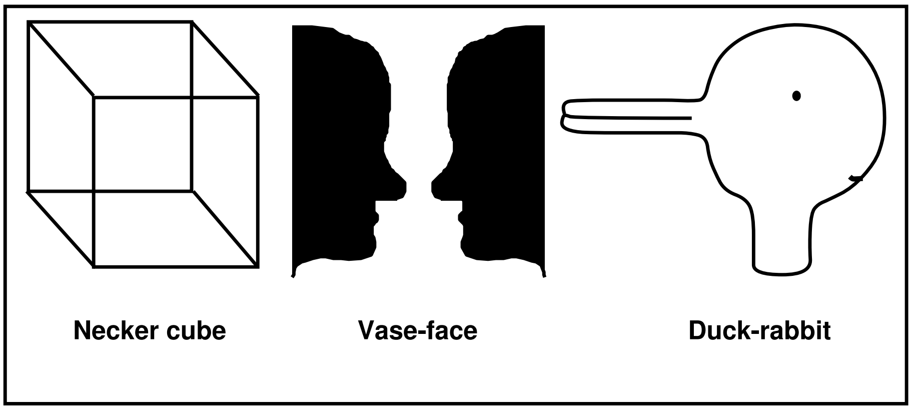
E.g. the towers may be thick or thin. They may have internal processing layers.

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

Levels in perceptual mechanisms

Detection of low level physical changes at transducers, detection of 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.

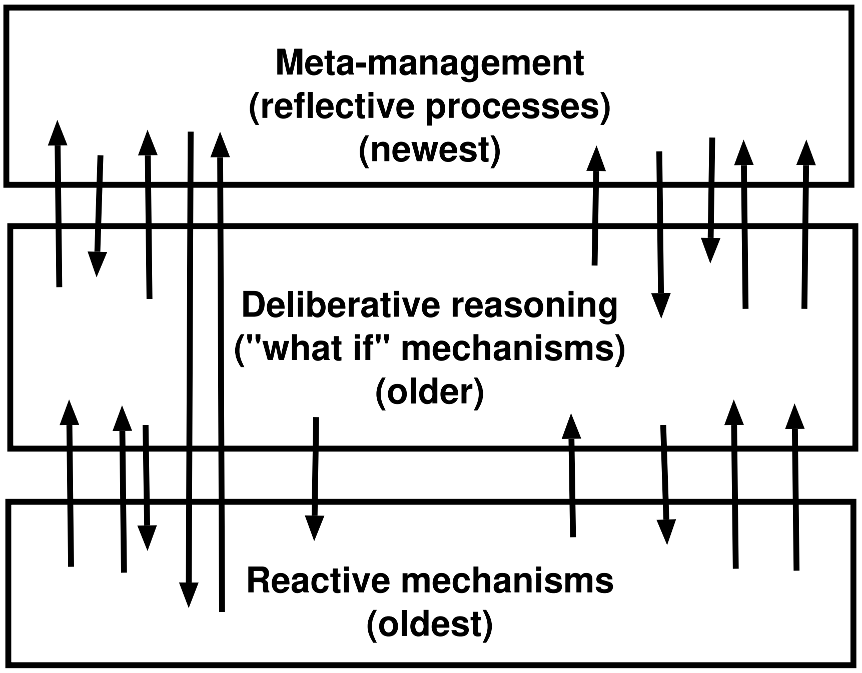
These in turn can evolve to make new uses of more sophisticated perceptual layers.

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.

ONE OF MANY LAYERED VIEWS



Different layers need not map onto different portions of the brain in any simple way.

Contrast: MacLean's theory of the triune brain: reptilian, old mammalian, new mammalian and the partially similar layered theory of James Albus in *Brains, Behaviour and Robotics* 1981.

Our layers do not form a dominance hierarchy.

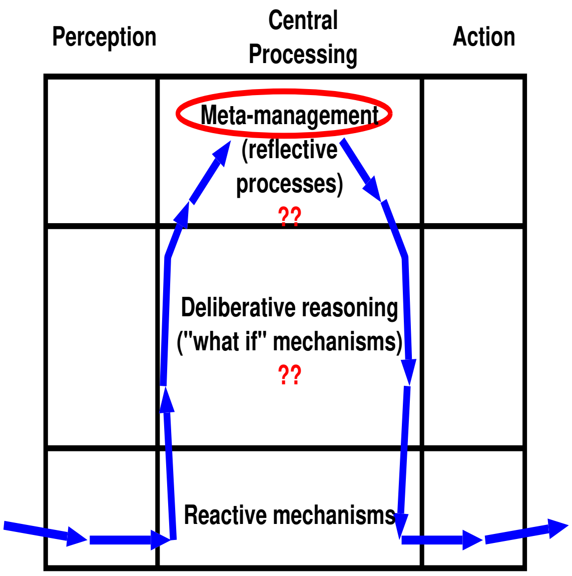
Layered architectures have many variants

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



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 recently elaborated the SAS: Supervisory Attention System at the 'top'. The ideas overlap with meta-management.

CogAff allows more ‘horizontal’ connections.

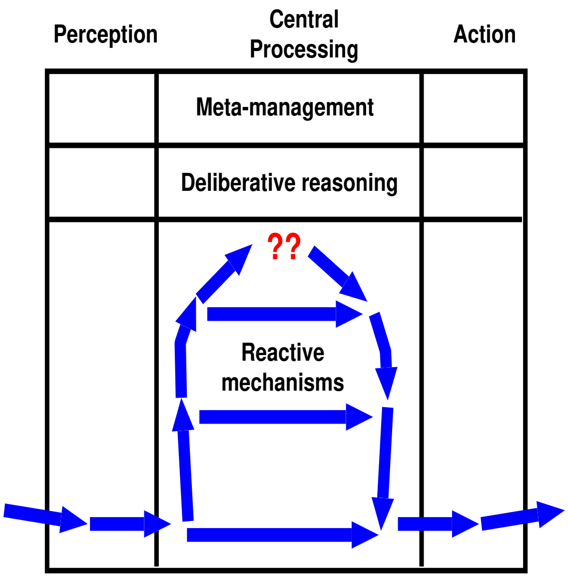
Most systems differ from the CogAff framework by not allowing the perception and action systems to include hierarchies of abstraction with direct connections at all levels to central layers. Hence the horizontal connections are only at the lowest level: ‘peephole’ vs ‘multi-window’ perception and action.

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)

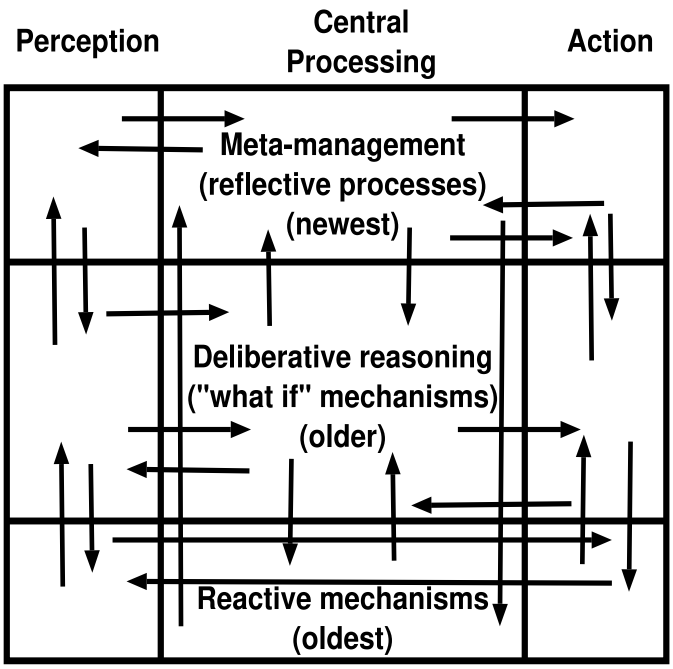


This allowes layers of control, within the reactive category, but Brooks (sometimes) denies that animals use deliberative mechanisms.

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

LAYERS + TOWERS = GRID

Of co-evolved concurrently active sub-organisms, each contributing to the "niches" of the others.



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.

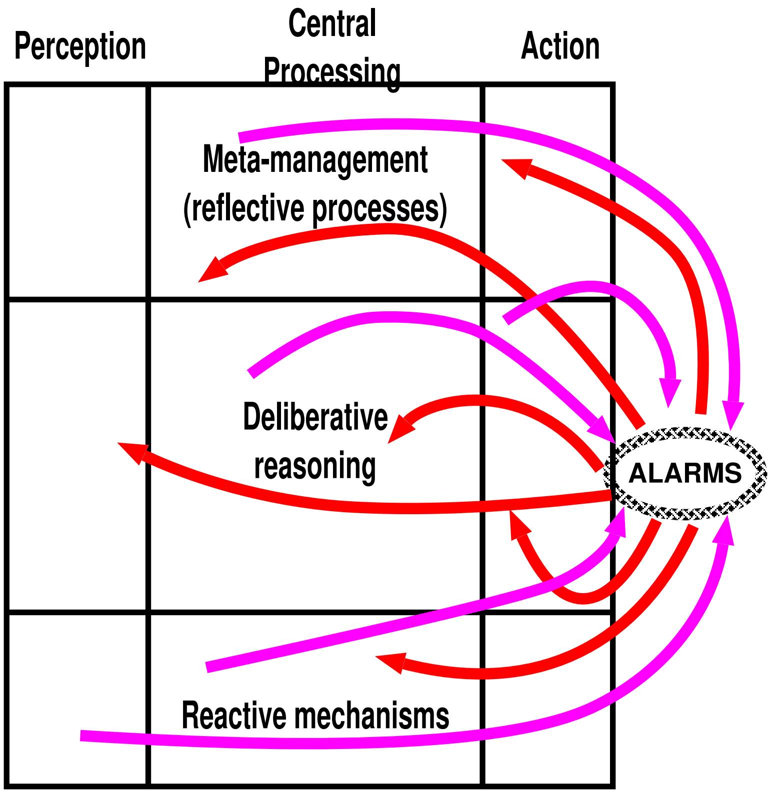
Then no one layer dominates the rest (as in subsumption)

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.

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

The alarms are drawn outside the grid for clarity, but require mechanisms that are in the reactive layer.

ADDITIONAL COMPONENTS

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

MANY PROFOUND IMPLICATIONS

e.g. for kinds of development
kinds of perceptual processes
kinds of brain damage
kinds of emotions
and other affective states

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.

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?

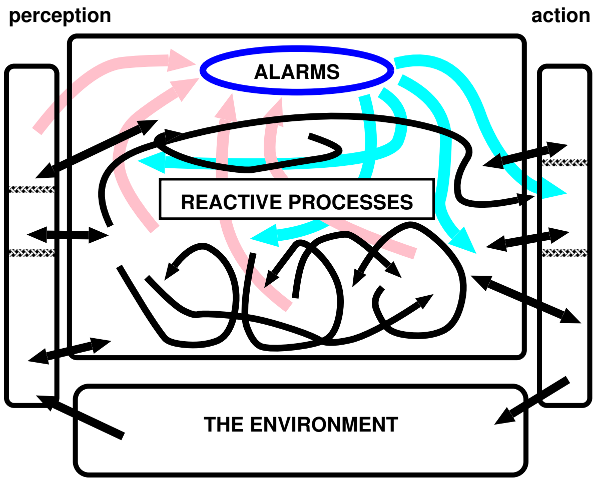
Can any insects do deliberation? Any fish? Any reptiles?

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?



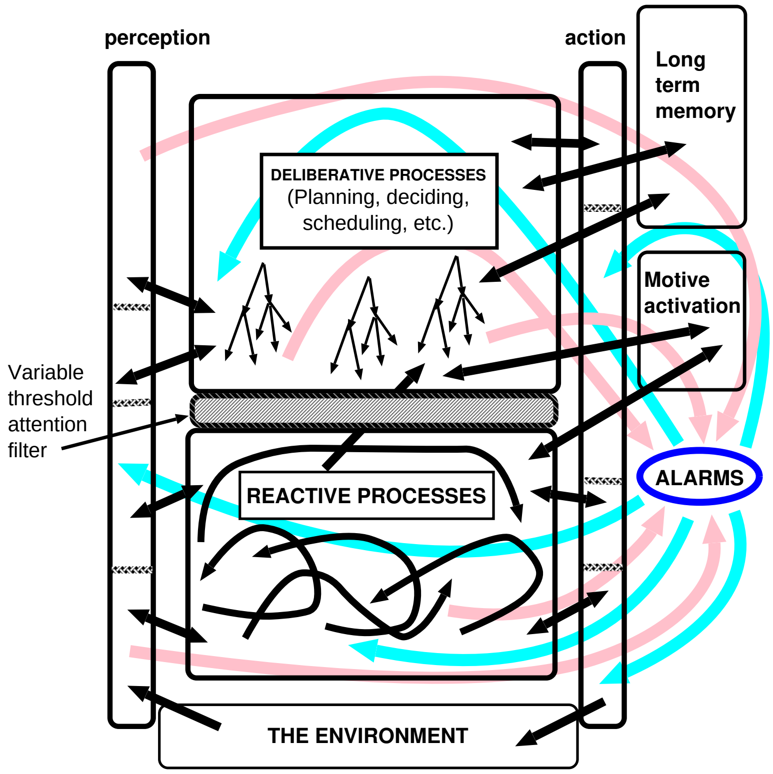
ALARM MECHANISM (Global interrupt/override):

- **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.**

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!

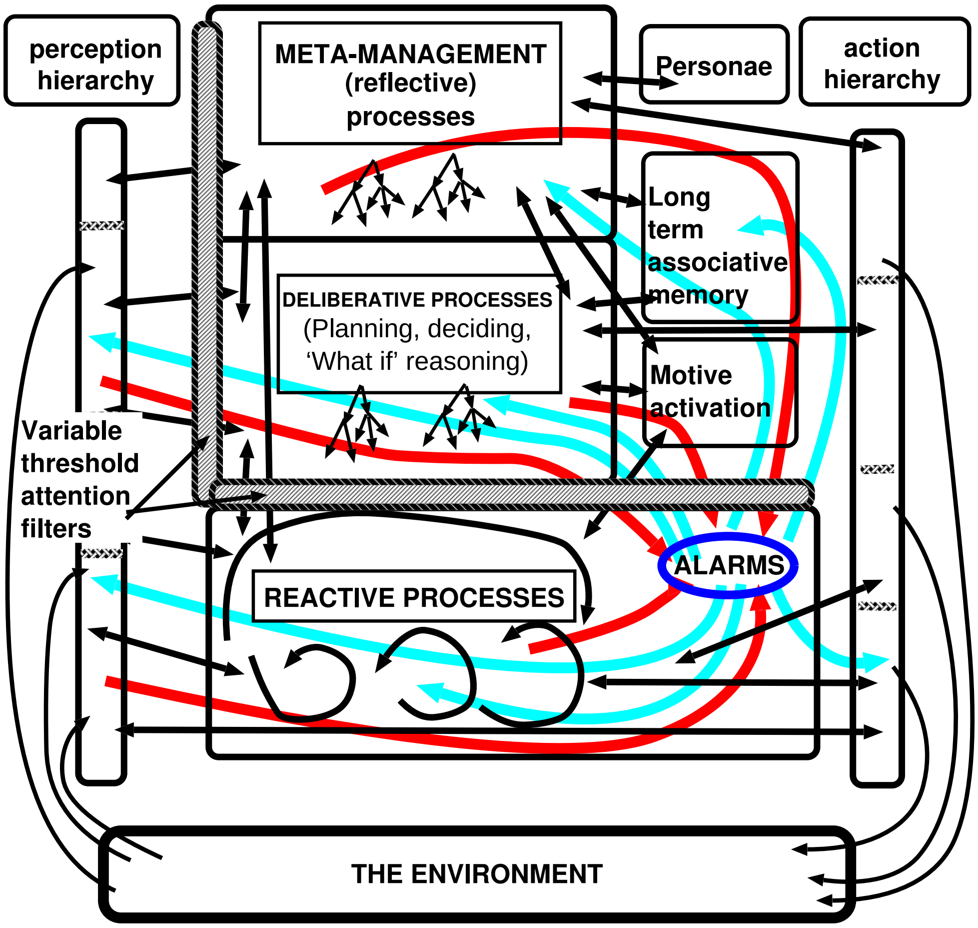
When we add the meta-management layer, we find scope for another class “tertiary emotions”.

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/

The H-Cogaff Architecture



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

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.

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)**

MAYBE WE SHOULD GET TOGETHER ON THIS...

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

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)

THESE AND RELATED SLIDES CAN BE FOUND IN

http://www.cs.bham.ac.uk/~axs/misc/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 subsystems,**
- **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.