Presented at Nokia Research Centre, Helsinki, 8th June 2001

VARIETIES OF EVOLVABLE MINDS

OR

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

These slides are at: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#nokia

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.

AND THANKS TO LINUX DEVELOPERS, AND OTHERS WHO ALLOW ME TO USE COMPUTERS WITH RELIABLE FREE SOFTWARE

MAIN THEMES

- We need to work with an enriched ontology:
- We need to understand how evolution produces such things.
- We need to investigate particular architectures in depth.
- We need to use the results to clarify our pre-scientific concepts, like “consciousness”, “learning”, “emotions”, etc.

MAIN THEMES

- **We need to work with an enriched ontology:**
  - A broader conception of “information processing” machines
  - Events and processes in virtual machines
  - Architectures for such machines
  - Causal powers of such machines
  - Multiple implementation/supervenience levels
- **We need to understand how evolution produces such things.**
- **We need to investigate particular architectures in depth.**
- **We need to use the results to clarify our pre-scientific concepts, like “consciousness”, “learning”, “emotions”, etc.**

MAIN THEMES

- **We need to work with an enriched ontology:**

  (As on previous slide)

- **We need to understand how evolution produces such things.**
  - Myriad varieties of biological information processing systems, on different scales: cells, organisms, colonies/societies, ecosystems, etc.

    These are produced by a huge variety of processes: evolution, individual developmental, learning, social interaction, and many patterns of co-evolution.

- **We need to investigate particular architectures in depth.**

- **We need to use the results to clarify our pre-scientific concepts, like "consciousness", "learning", "emotions", etc.**

MAIN THEMES

- **We need to work with an enriched ontology:**
- **We need to understand how evolution produces such things.**
- **We need to investigate particular architectures in depth.**

  This requires a conceptual framework for investigating architectures, especially information-processing architectures, especially virtual machine architectures, e.g.

  – the CogAff schema for describing designs within part of the space, (discussed below)
  – the H-Cogaff architecture for human-like cases, (also discussed below)
  – other architectures, e.g. for insect colonies, human societies, and other multi-agent systems, (not discussed below!).

- **We need to use the results to clarify our pre-scientific concepts, like "consciousness", "learning", "emotions", etc.**

MAIN THEMES

- **We need to work with an enriched ontology:**
- **We need to understand how evolution produces such things.**
- **We need to investigate particular architectures in depth.**
- **We need to use the results to clarify our pre-scientific concepts, like "consciousness", "learning", "emotions", etc.**
  - Analyse the conceptual confusions in these concepts
  - Show how new architecture-based concepts can provide far greater conceptual clarity and precision.
  - Many of our intuitive ideas that seem obviously true are just cultural or individual prejudices, and will have to be rejected or displaced: e.g. ideas about other animals, about new born babies, about requirements for consciousness, about functions of consciousness, about what free will is, etc.

MAIN THEMES

- **We need to work with an enriched ontology:**
- **We need to understand how evolution produces such things.**
- **We need to investigate particular architectures in depth.**
- **We need to use the results to clarify our pre-scientific concepts, like "consciousness", "learning", "emotions", etc.**

The results of this project will not only revolutionise philosophy of mind, psychology and neuroscience, it will also be very relevant to many engineering possibilities.

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.

Key idea: they are information processing machines

WHAT IS INFORMATION?

The concept of “information” is partly like the concept “energy”.

It is hard to define “energy” in a completely general way.

Did Newton understand what energy was?

There are many kinds that had not yet been conceived of.

We can best think of energy in terms of:

- the different forms it can take,
- the ways in which it can be transformed, stored, transmitted, or used, • the kinds of causes and effects that energy transformations have,
- the many different kinds of machines that can manipulate energy
- ....

Science does not start with definitions: implicit definitions emerge as theories develop.

See also
http://www.cs.bham.ac.uk/research/cogaff/talks/#inf

We do not need an explicit definition of "energy"

It is a primitive theoretical term – implicitly defined by the processes and relationships that involve it.

We should not use currently known forms of energy to *define* it, since new forms of energy may turn up in future.

Newton knew about energy, but did not know anything about the energy in mass. The relation $E = MC^2$ had not been thought of.

We may discover yet more new forms of energy and new forms of energy transformation.

Likewise, our understanding of "information" and "information processing" is currently partial.

What is information?

A full theory would specify:

- the different types of information,
- the different forms in which they can be expressed,
- the different ways information can be acquired, transformed, stored, searched, transmitted or used,
- the kinds of causes that produce events involving information,
- the kinds of effects information states and information changes can have,
- the many different kinds of machines that can manipulate information,
- the variety of *architectures* into which information processing mechanisms can be combined

If we understand all that, we don't need to define "information"!

Like "energy", "information" is an implicitly defined primitive theoretical term that plays a role in explanatory theories.

We can explain many things that animals and machines do in terms of the information they have access to. Plants too.

A Flawed Common Assumption

The common assumption that "information processing" necessarily involves computers, especially computers as we know them, is based on limited imagination.

WE ARE IN THE VERY EARLY STAGES OF EXPLORATION, OF

- **INFORMATION PROCESSING MECHANISMS,**
- **ARCHITECTURES,**
- **APPLICATIONS.**

Compare physics at the time of Galileo?

Examples of types of processes involving information

- Acquiring
- Filtering/selecting
- Transforming/interpreting/disambiguating
- Compressing/generalising/abstracting
- Deriving (making inferences, but not only using propositions)
- Storing/Retrieving (many forms: exact, pattern-based, fuzzy)
- Training, adaptation (e.g. modifying weights, inducing rules)
- Constructing (e.g. descriptions of new situations or actions)
- Comparing and describing information (meta-information)
- Combining different items of information in new ways
- Reorganising (e.g. formation of new ontologies)
- Testing/interrogating (is X in Y? is A above B? did C cause D? what's the P of Q?)
- Exporing, considering, hypothesising
- Syntactic manipulation of information-bearing structures
- Translating between forms, e.g. propositions, diagrams, weights
- Controlling/triggering/modulating behaviour (internal, external)
- Propagating (e.g. in a semantic net, or neural net)
- Transmitting/communicating

NOTE: A machine or organism may do some of these things internally, some externally, and some in cooperation with others.

The processes may be discrete or continuous (digital or analog).

Differences between information and energy

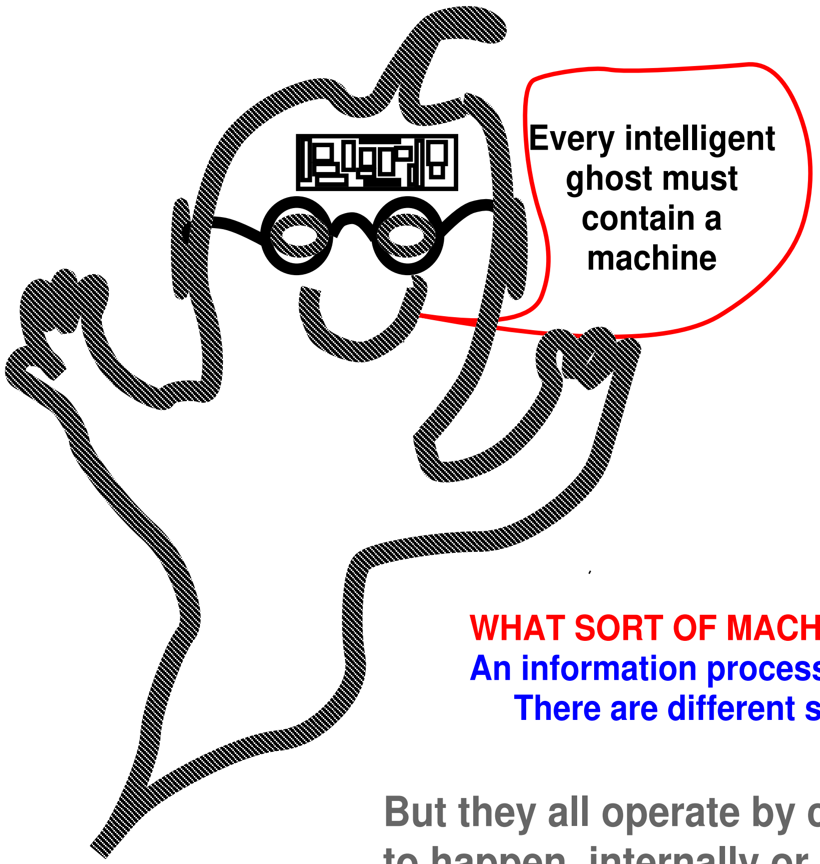
It is very useful to *measure* energy e.g. because it is conserved.
But measuring information is often less useful.

- I give you information, yet I still have it, unlike energy.
- You can derive new information from old, and still have both.
- Information varies primarily not in its *amount*, like energy, but in its structure and content.
- Numbers (measurements) do not capture what is most important about information, for behaving systems: namely structure, content, applicability.
- Equations do not represent most information manipulations adequately: Compare programs, proofs, and other forms of reasoning.

However, like energy, information can be involved in complex interacting processes, which involve not only changes in information, but also physical effects.

Do we need a ghost in the machine?

What philosophers tend to forget - but the ghost of Gilbert Ryle knows well....



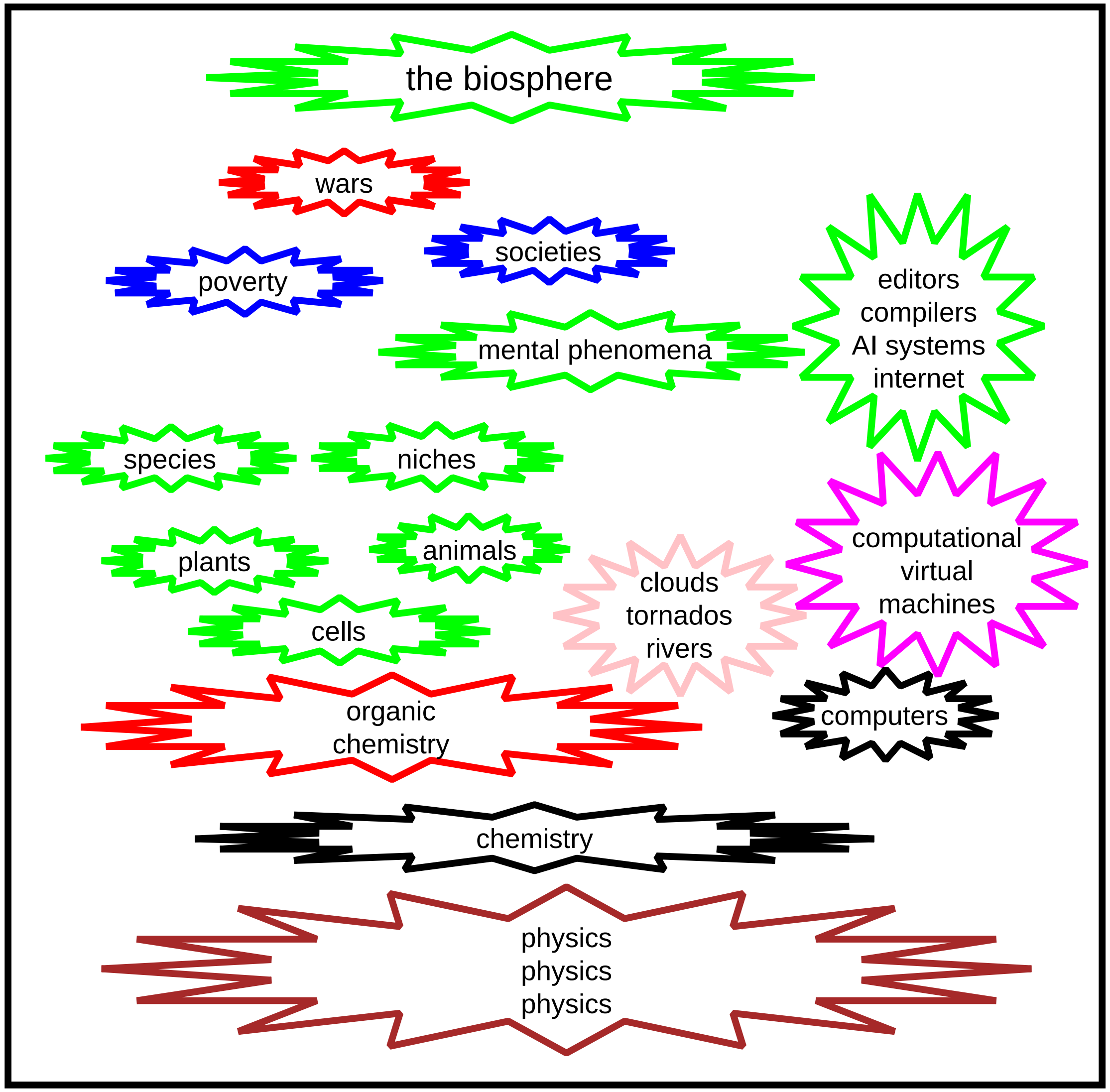
WHAT SORT OF MACHINE?

An information processing machine.
There are different sorts.

But they all operate by causing things to happen, internally or externally.

Every virtual machine 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?

ENGINEERS AS PHILOSOPHERS

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", well understood intuitively by software engineers.

(Though they may not be able to articulate what they understand).

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

What engineers implicitly understand, however is very complex, and hard to make precise.

There are different sorts of supervenience (realisation):

- **Property supervenience,**
  E.g. the property of a painting of being in the style of Picasso supervenes on lower level properties of parts: there cannot be two paintings P1 in the style of Picasso, P2 not, unless P1 and P2 also have some physical difference.
- **Pattern supervenience,**
  E.g. various patterns of rows, columns, diagonals, squares, etc. can supervene on the same physical arrangement of dots.
- **Agglomerative, or mereological (part/whole), supervenience,**
  E.g. angular momentum of a wheel, supervenes on properties of the atoms.
- **Mechanism supervenience:**
  An ontology which includes many concurrent processes and causal interactions is implemented in a lower level ontology.

We understand only a tiny subset of the space of possible virtual machine architectures, and the ways they can supervene on physical infrastructure.

We need to understand the space, to understand what minds are, and how they can exist.

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 of minds will not come from studying ONE variety – e.g. typical adult human minds!

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”.

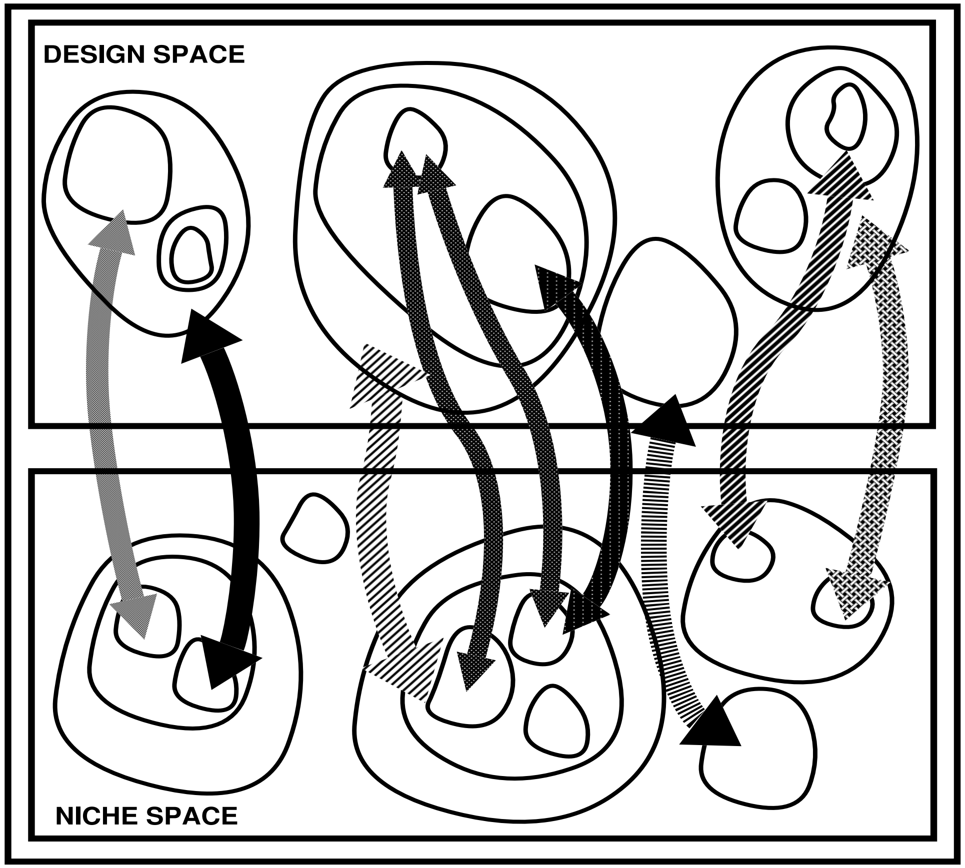
A possible exception: evolution can use the cognitive abilities of “intelligent” informed individuals, e.g. in mate selection.

Evolution produces *niches* as well as *designs*

A design or a niche is produced by a process. We need to understand the variety of processes and the designs and niches they can produce.

A process involves a trajectory of some system through a space.

DESIGN SPACE AND NICHE SPACE



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

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

Different sorts of trajectories through the 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.

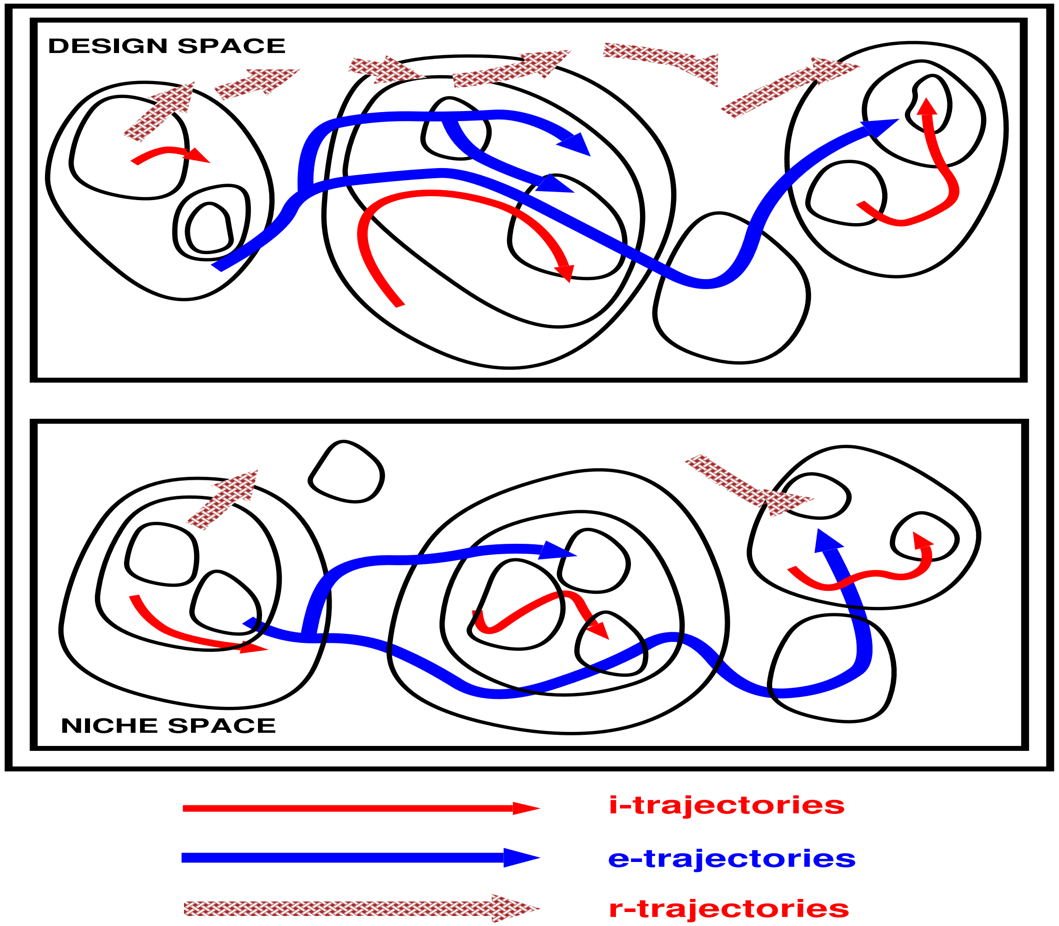
s-trajectory: possible for social systems with multiple communicating individuals. (Can be viewed as a type of i-trajectory.)

c-trajectory: one of the above which makes use of the cognitive abilities of what is developing or evolving.

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

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

Varieties of trajectories



An external "repairer" can push something through an "r-trajectory" in which intermediate forms are not all viable.

Biological evolution is discontinuous: e.g. mutation, crossover, etc.

Biological evolution

- Multiple interacting **e-trajectories**,
- later using **i-trajectories**,
- then **s-trajectories**,
- and now also **r-trajectories** (e.g. genetic engineering).

When cognitive mechanisms evolve they can support

- **“c-trajectories”**, using the *cognitive* abilities of individuals to modify e-trajectories.

Many unanswered questions:

– 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?

See papers by Chappell and Sloman on the altricial precocial spectrum for animals and machines.
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0502
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609

EVOLUTION OF MIND: Different mental concepts are applicable in different architectures

- An architecture supports a collection of possible states, processes, causal interactions:
- Different collections are possible in different architectures.
- If mental concepts are architecture-based then we can't apply the same concepts (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

  The fly detects the approaching object and reacts by buzzing off.

  But it probably does not know what it has done or why it has done it.

  So it has a kind of consciousness but not our second-order self-awareness, although it may have a primitive kind of self-awareness, e.g. in detecting when it needs to feed.

- 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.

Apparently similar animals may have very 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...)
- When adults of the two types look at the same scene, they are almost certainly processing it very differently, e.g. deer vs lion.

  The deer's visual system was designed by evolution.
  Some aspects of the lion's visual archtecture were 'designed' and grown by the lion as it interacted with the environment.
  Compare humans who learn to read different languages.

For more on this see the papers by Sloman and Chappell here
http://www.cs.bham.ac.uk/research/projects/cosy/papers/

E.g. http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609
COSY-TR-0609: Altricial Self-organising Information-processing systems

Why are precocial and altricial species so different?

To answer, we need to compare the design requirements (niches) for adults of both types.

- **Deer, sheep, horses, etc. need to be able to feed, walk and run from predators on grassy plains.**
- **Compare the needs of adult chimps, monkeys, leopards, eagles, squirrels.**
- **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, self-programming architecture.**
- **Extended development times support more sophisticated self-programming.**
- **So we probably need different sets of concepts to describe what an adult lion sees and what an adult 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.

You may not have the concepts to describe the experiences of a human from another culture.

Ignoring that is another variety of “anthropomorphism”!

Even within a culture, a mathematician’s mind could have a different architecture from a dancer’s, or a chimney sweep’s.

Mental concepts are ‘cluster’ concepts

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

- different kinds of learning — MANY kinds
- many variants 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 and other affective states

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

This makes the question hard to answer.

Perhaps there is no question to answer?

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 – but there are probably a lot more than three kinds.

Architectures vs Representations & Algorithms

In the 1960s and 1970s AI used to be mainly about representations and algorithms.
Since the mid-1980s questions about architectures have been widely recognised as 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 layers which evolved at different times and use different mechanisms:

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

We can also divide the functionality:

- SENSORY/PERCEPTUAL SYSTEMS
- INTERNAL PROCESSING
- MOTOR SYSTEMS

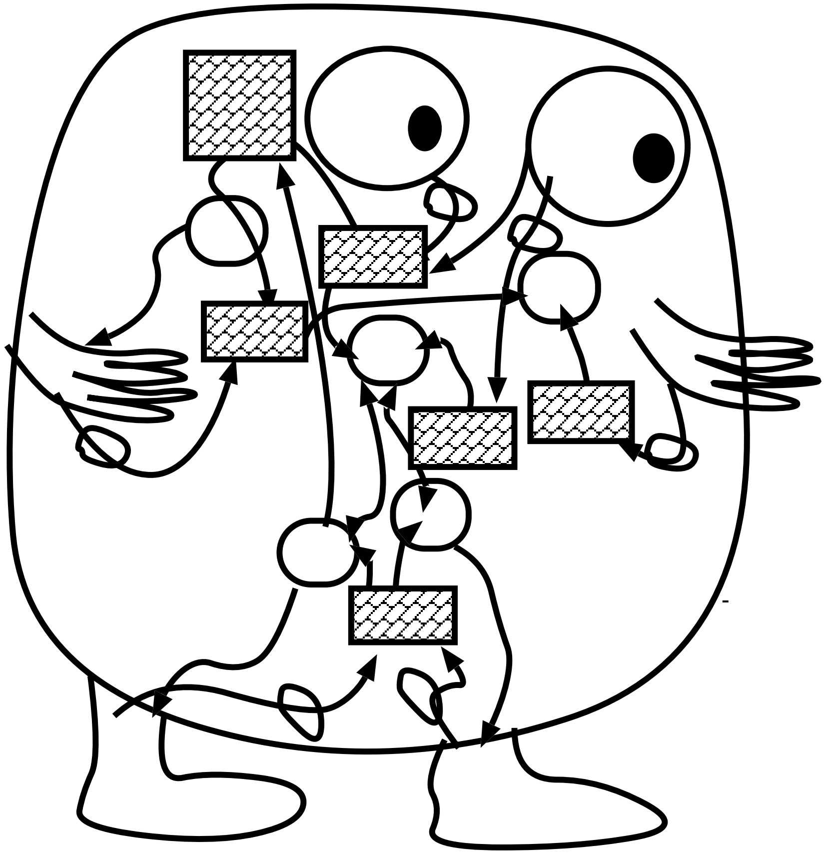
We need some good organising ideas.

- 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".

Do we know what sorts of architectures can be produced by evolution in different contexts?

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 (sketched later) 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.

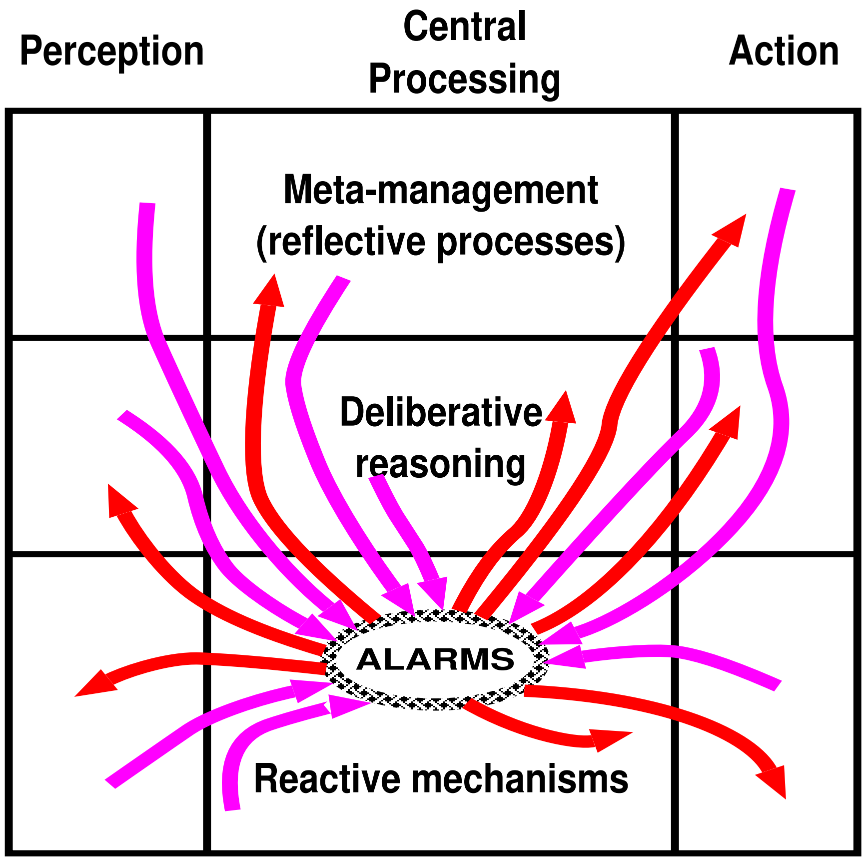
- various kinds of psychiatric disorder,
- types of self-delusion,
- limitations of short-term memory or processing accuracy,
- buggy interrupt systems (leading to some undesirable emotions),
- many kinds of fallacious reasoning
- religious beliefs,
- nationalism,
- racism,
- overconfidence in one's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

It is impossible to eliminate bugs in complex systems.

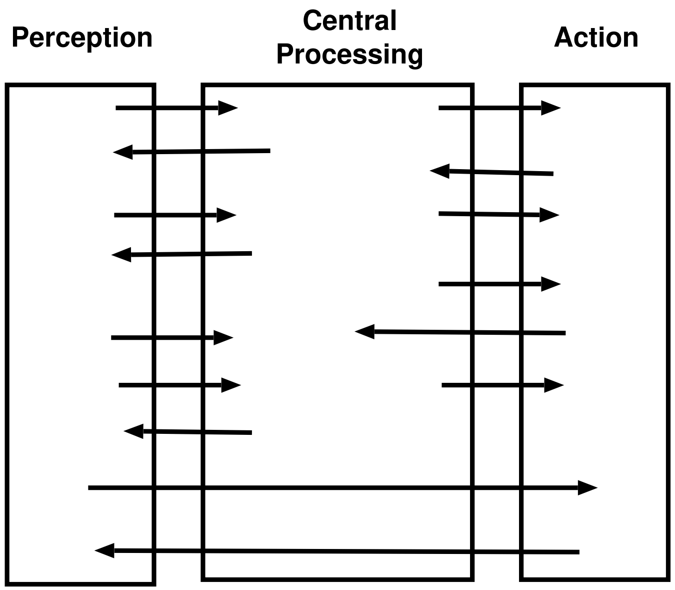
We need theories help to explain why some are likely.

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. Missing additional components are described later.



The “triple tower” View



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

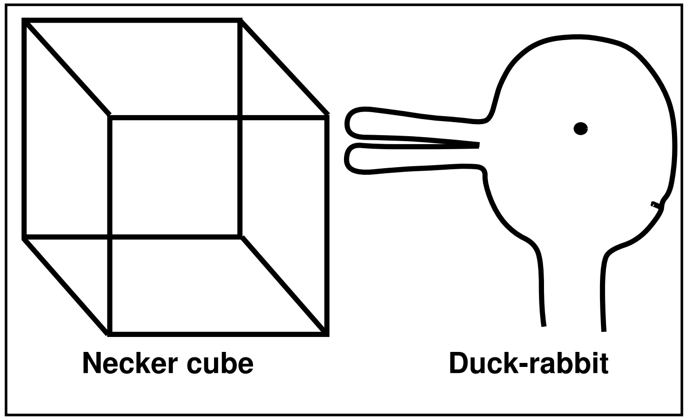
Many variants: (NILSSON, ALBUS)

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

E.g. 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!**

Extending Gibson's theory: Evolution of perceptual mechanisms

Different perceptual sub-systems recognize different affordances, using different ontologies to do so.

LIKE DIFFERENT ORGANISMS

Different levels of perceptual abstraction required for different purposes. E.g. "chunking" required for learning correlations.

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.

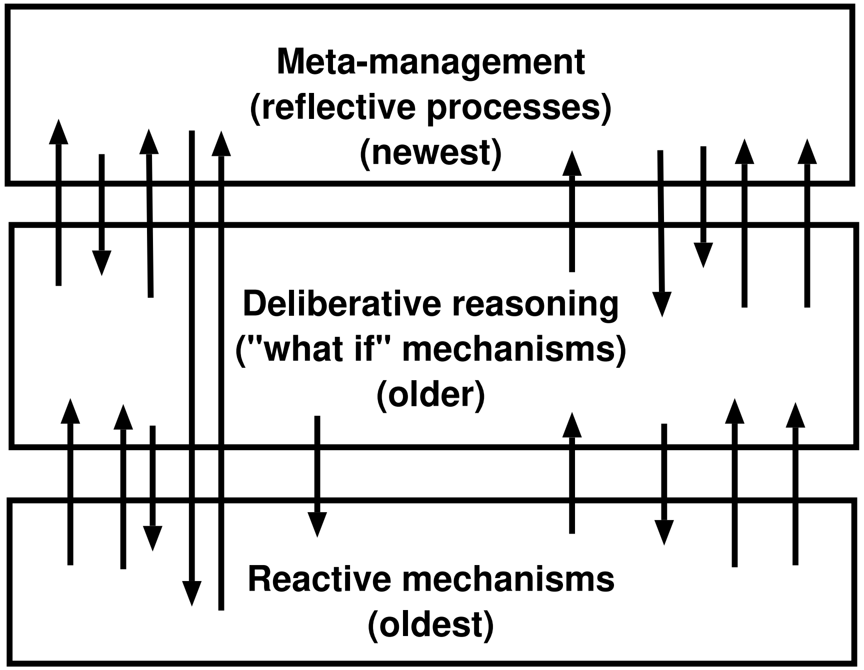
Likewise layered action systems.

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



Compare the “triune” brain: reptilian, old mammalian, new mammalian.

How do the layers differ in their mechanisms, and capabilities?

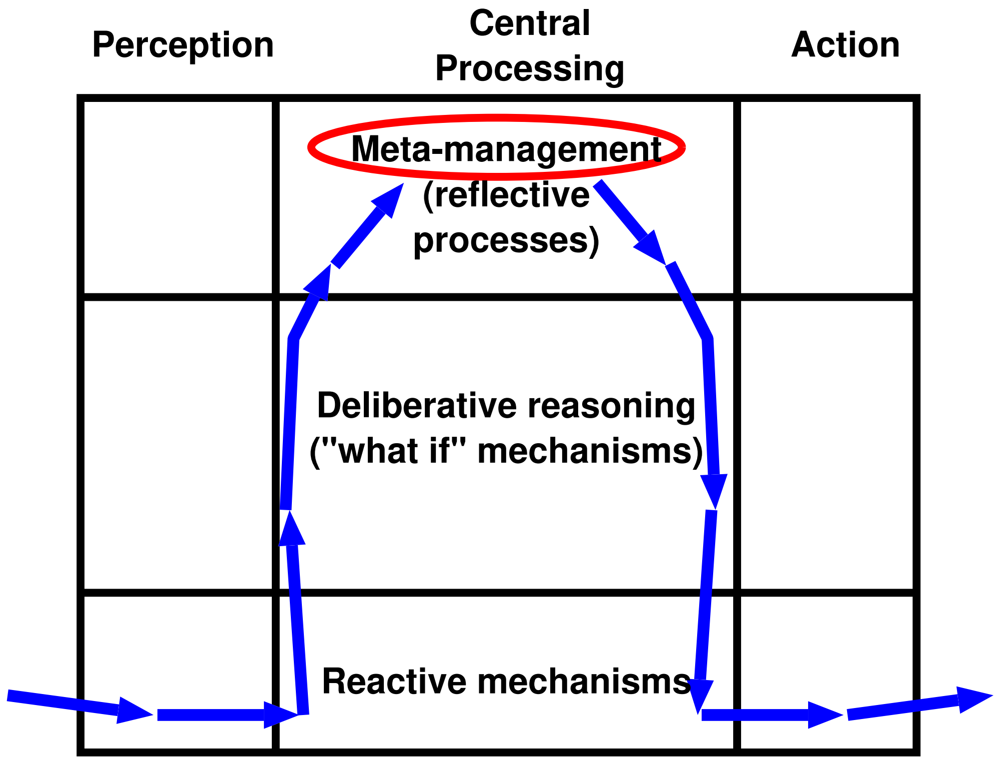
Layered architectures have many variants

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

Different principles of subdivision in layered architectures

- evolutionary stages (reactive earliesr)
- levels of abstraction
  (e.g. to support "what if" reasoning in deliberative mechanisms, or to categorise one's own internal virtual machine proesses).
- 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.

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

Essentially a pipelined architecture.

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

SOME DIFFERENCES BETWEEN THE LAYERS

Reactive systems can be highly parallel, very fast, and use analog circuits.

Deliberative mechanisms provide ‘What if’ reasoning and representation capabilities.

For a more detailed analysys of requirements for ‘fully deliberative’ architectures, and comparisons with simpler intermediate cases see

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604
COSY-DP-0604: Requirements for a Fully Deliberative Architecture
(Or component of an architecture)

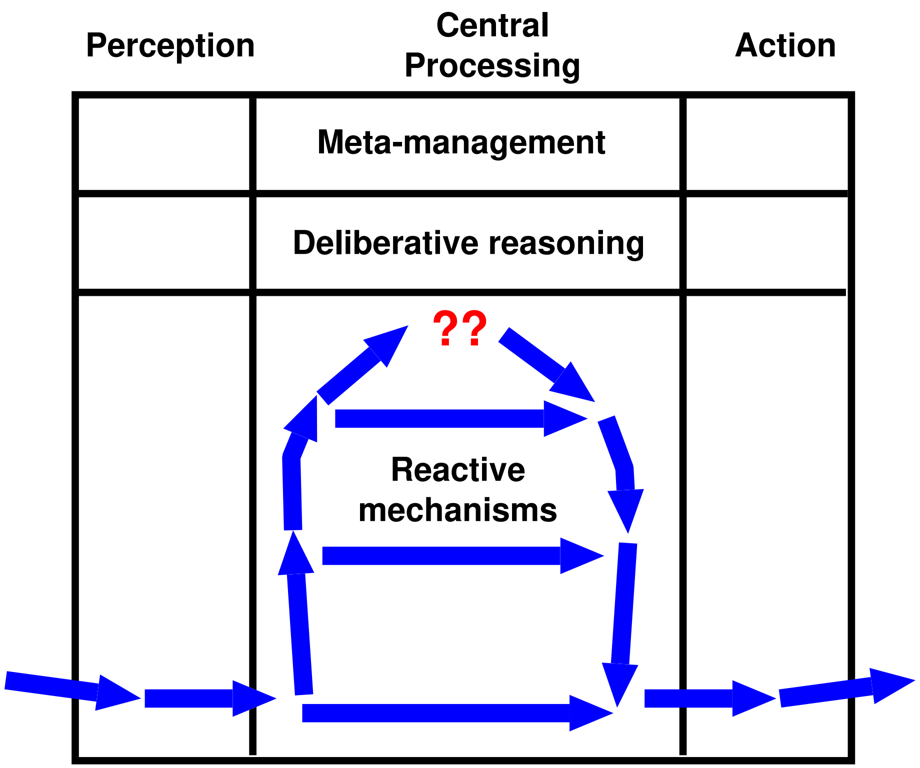
The most sophisticated deliberative mechanisms are inherently slow, serial, knowledge-based, resource limited.

(Why?)

Meta-management adds reflective abilities: this can sometimes speed things up through discovery of short cuts, bugs in thinking, etc.

Compare “executive function” in clinical usage.
(Combines/confuses the top two layers.)

Another variant: Subsumption architectures (Brooks, MIT)



Denies that any animals, including humans, use deliberative mechanisms.

(How do they get to overseas conferences?)

LAYERS + TOWERS = GRID

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

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

Many varieties of information flow and control flow.

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.

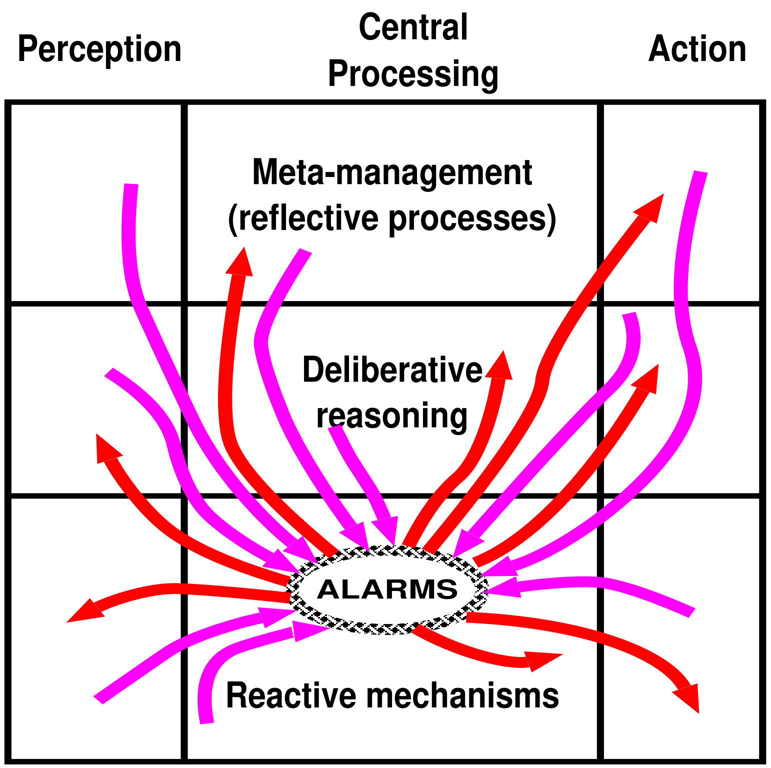
Then no one layer dominates the rest (as in subsumption)

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

"ALARM" MECHANISMS

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

Fast, powerful, "alarm systems" needed



“ALARM” MECHANISMS WILL BE STUPID

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).**

Many variants of the Schema possible.

E.g. one alarm system or several?

Brain stem, amygdala, various reflexes, high trained reactions, etc.

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

- for kinds of development
- for kinds of perceptual processes
- for kinds of brain damage
- for kinds of emotions
- for other affective states (desires, moods, attitudes, etc.)

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.

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.

A general schema for evolvable architectures should allow for different sorts of representations (formalisms, languages) serving different purposes in different parts of the architecture.

It is just a prejudice that language is for external communication.

See:

http://www.cs.bham.ac.uk/research/cogaff/sloman.primacy.inner.language.pdf
http://www.cs.bham.ac.uk/research/cogaff/Aaron.Sloman_towards.th.rep.pdf

Both part of the CogAff web site: http://www.cs.bham.ac.uk/research/projects/cogaff/

CogAff is a scheme not an architecture:

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?

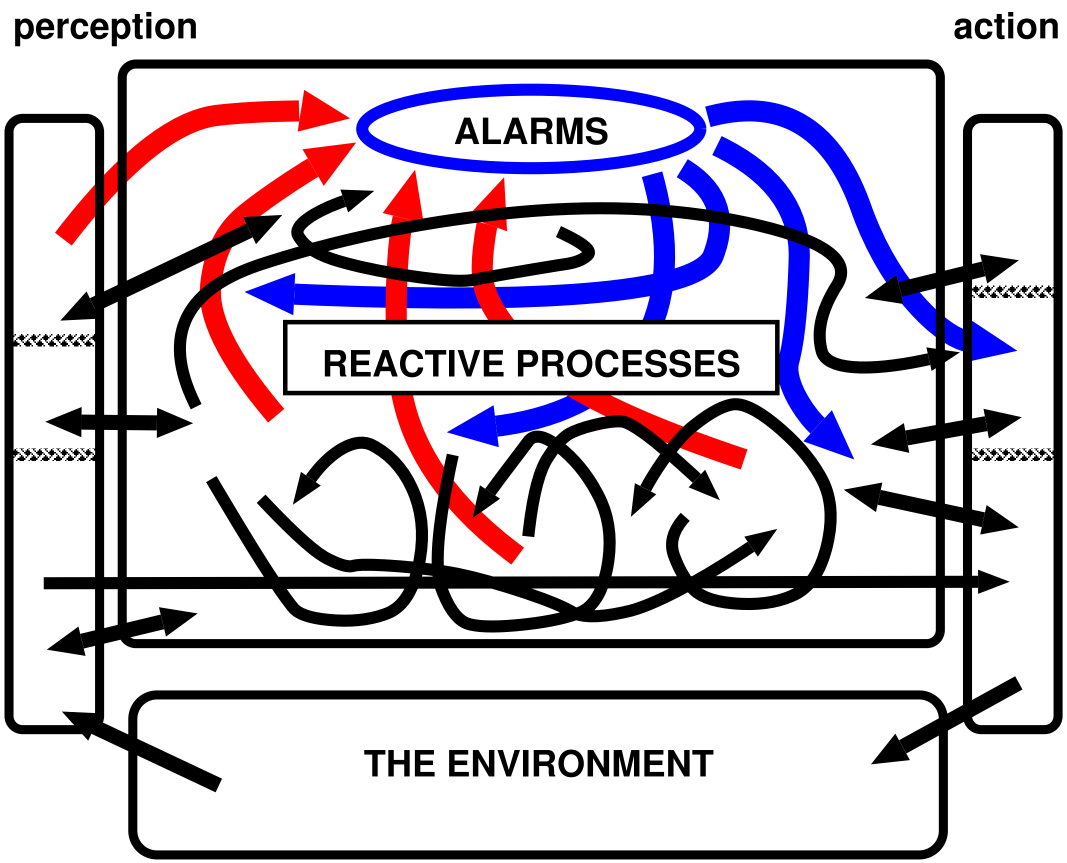
Can any insects do deliberation? Any fish? Any reptiles?

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?



Insects seem to lack a deliberative layer. They may have alarm mechanisms as part of ther reactive architecture.

(Reactive does not mean "stateless". It means "non-deliberative".)

ALARM MECHANISM (Global interrupt/override)

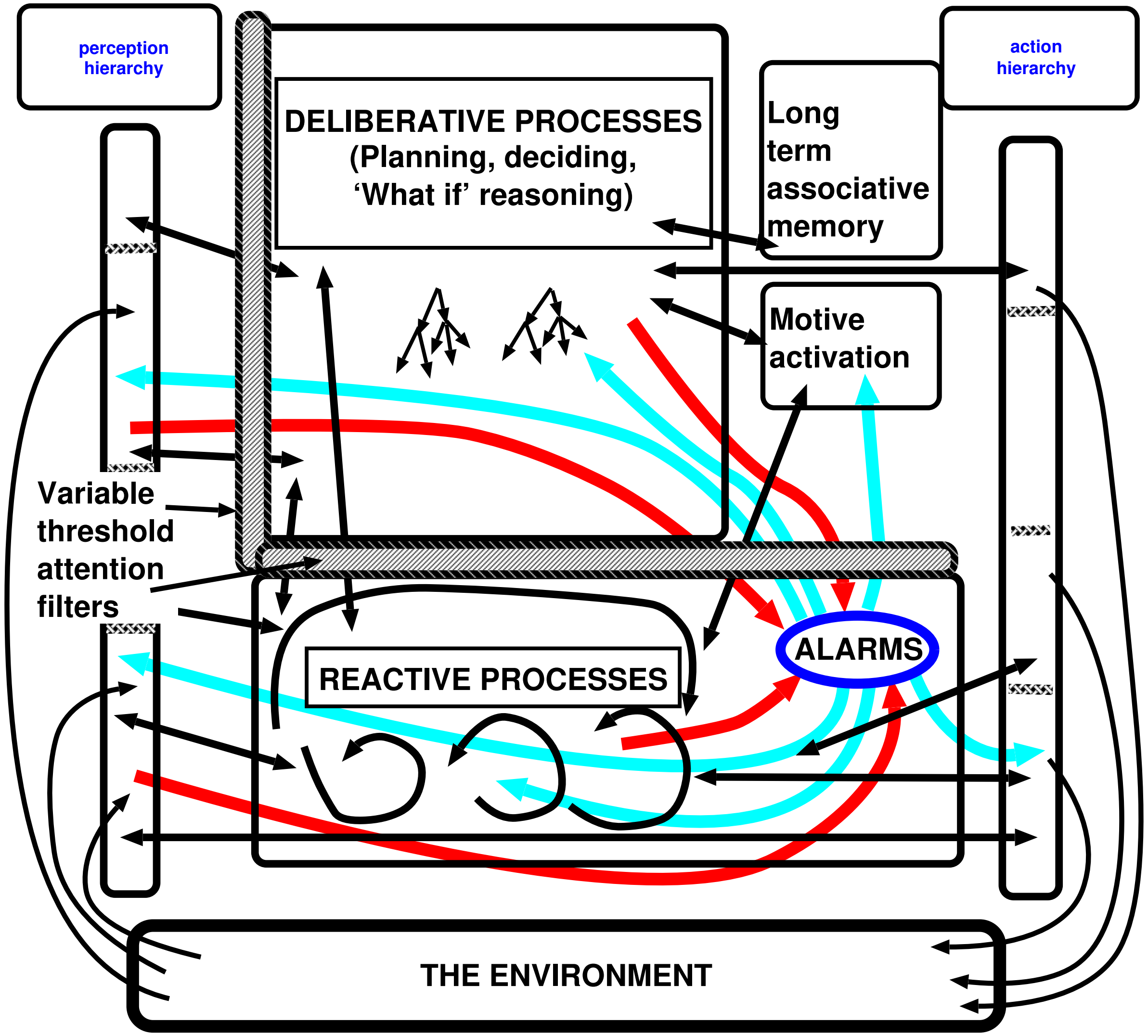
- Allows rapid redirection of the whole system, for sudden dangers or sudden opportunities
- FREEZING
- FIGHTING, ATTACKING
- FEEDING (POUNCING)
- GENERAL AROUSAL AND ALERTNESS (ATTENDING, VIGILANCE)
- FLEEING
- MATING
- MORE SPECIFIC TRAINED AND INNATE AUTOMATIC RESPONSES

Even simple insect-like organisms may need this sort of architecture – though for some animals it does not suffice.

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

Next figure illustrates comblined deliberative and reactive architectures.

Reactive and deliberative layers with alarms



HYBRID ARCHITECTURES IN NATURE

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

- **How many varieties of long term memory**
  SUPPORTING DIFFERENT KINDS OF DELIBERATION
- **How many 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)

ALARM MECHANISM
IN A HYBRID SYSTEMS

(BRAIN STEM, LIMBIC SYSTEM? DIFFERENT PATHWAYS?)

ALLOWS RAPID REDIRECTION OF THE WHOLE SYSTEM.

But can be triggered by and can redirect deliberative processes, as well as reactive 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 hybrid architectures

Damasio & Picard: Cognitive processes can 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".

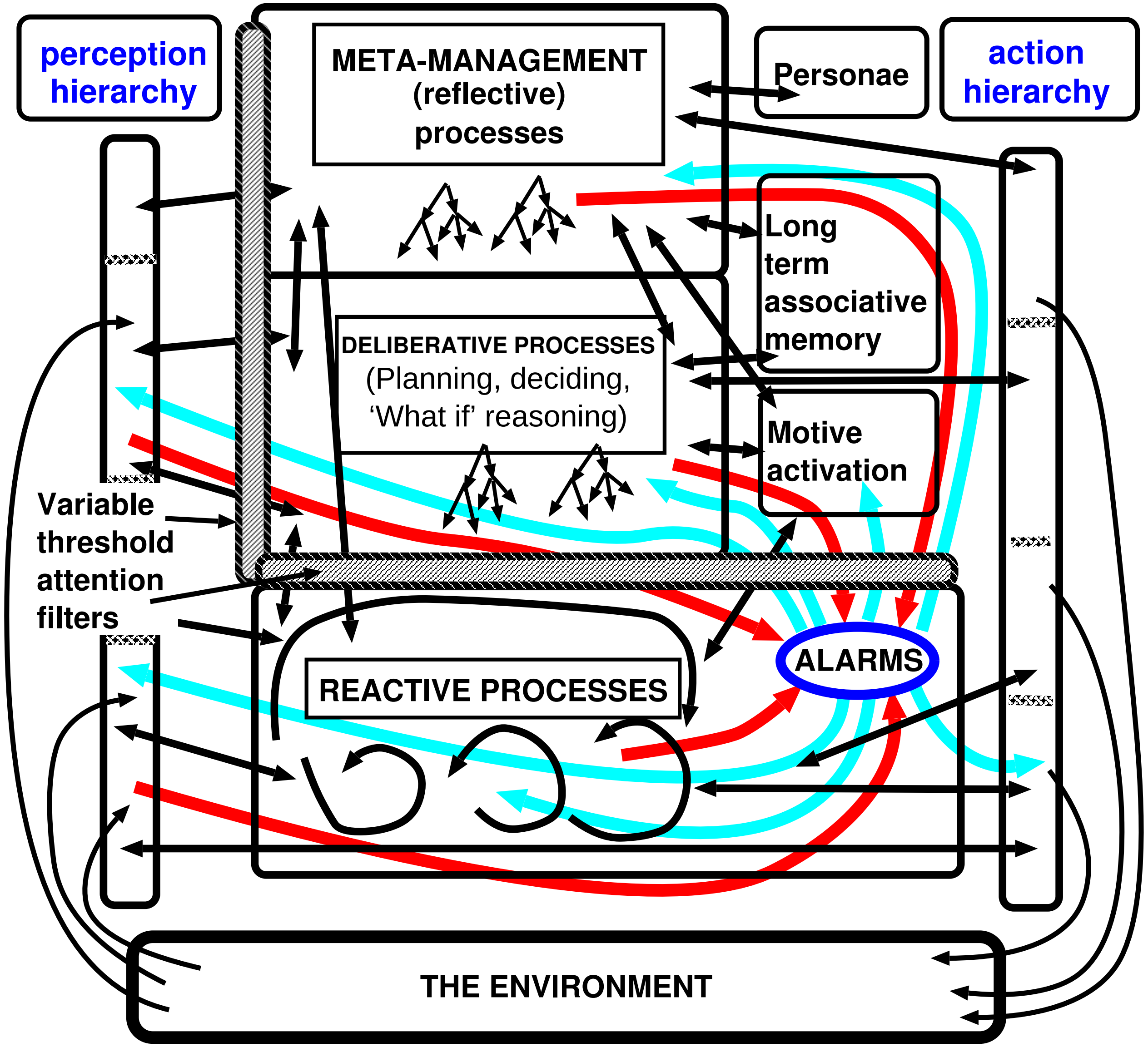
Thinking about too narrow a range of architectures (or not thinking about architectures) can hamper the search for explanatory theories.

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

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

The H-Cogaff Architecture: an instance of CogAff

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

Explaining disputes and conflicting definitions

E.g. there are many different, and inconsistent, definitions of "emotion" "learning" "executive function" etc.

Why?

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 there 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:

– if they want to model human affective processes,

– if they wish to design systems which engage fruitfully with human affective processes,

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

– 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

Also the CoSy project web site:

http://www.cs.bham.ac.uk/research/projects/cosy/papers/

THE END
(for now)