Biotechnology & Biological Sciences Research Council
Engineering & Physical Sciences Research Council

ADAPTIVE AND INTERACTIVE BEHAVIOUR OF ANIMALS AND COMPUTATIONAL SYSTEMS (AIBACS):
A Joint EPSRC/BBSRC Initiative.

Workshop at Cosener's House, Abingdon
28-29th March 2001

How to understand natural minds of many kinds.

A talk presented at the above workshop, on 28th March 2001.

AARON SLOMAN

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

COGNITION AND AFFECT PROJECT

SCHOOL OF COMPUTER SCIENCE

THE UNIVERSITY OF BIRMINGHAM

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

Aim of the joint research council initiative on Adaptive and Interactive Behaviour of Animals and Computational Systems (AIBACS)

EXTRACT FROM DOCUMENT CIRCULATED IN ADVANCE OF THE MEETING IN MARCH 2001:

> **"Aim: To improve our understanding of the mechanisms and organisational principles involved in the generation of adaptive and interactive behaviour in animals (including humans) and by computational systems.**
>
> **This can best be achieved through collaboration between neuroscientists (including the study of human psychology and animal behaviour), social scientists, computer scientists and engineers."**

Add philosophers too?

Biological systems and computing systems are both concerned with processing of information.

At present they generally process different sorts of information and in different ways, about which our understanding is still very limited.

Proposal: The AIBACS initiative should include investigation of the variety of types of information processing and how biological information processing architectures evolve and develop.

BIOLOGICAL INFORMATION PROCESSING

Biologists are used to thinking of genes as carrying information, and reproduction as transfer of information.

But scientists increasingly realise that most (or all) biological processes, including perception, learning, choosing, and behaving, involve the processing and use of information.

There are different kinds of information, for instance:

- about categories of things (big, small, red, blue, prey, predator)
- about generalisations (big things are harder to pick up)
- about particular things (that thing is heavy)
- about priorities (it is better to X than to Y)
- about what to do (run! fight! freeze! look! attend! decide now!)
- about how to do things (find a tree, jump onto it, climb...)

This categorisation of types of information does not cover all the types found in machines and organisms. We probably still know only about a small subset of types of information, types of encoding, and types of uses of information.

Don't expect all types to be expressible in languages we can understand – e.g. what a fly sees!

WHAT IS INFORMATION?
The concept is obscure: compare "energy"

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 he did not know about.

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
- ....

If we understand all that, then we don't need to *define* "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: $E = MC^2$ had not been thought of.

What is information?

Likewise we need to understand:

- **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 manipulation 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, then we don't need to *define* "information"!

Like "energy" it is an implicitly defined primitive theoretical term. Contrast metaphorical terms.

One big difference: 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. Equations do not represent most information manipulations adequately.**

Numbers (measurements) do not capture what is most important about information, for behaving systems.

EXAMPLES OF TYPES OF PROCESSES INVOLVING INFORMATION

- **Acquisition**
- **Filtering/selecting**
- **Transforming/interpreting/disambiguating**
- **Compressing/generalising/abstracting**
- **Deriving (making inferences, but not only using propositions)**
- **Storage/Retrieval (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)**
- **Reorganising (e.g. formation of new ontologies)**
- **Testing/interrogating (is X in Y, is A above B, what is the P of Q?)**
- **Copying/replicating**
- **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**
- **.... (many more)**

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).

Requirements for Information Processing

Not all the processes listed previously are possible in all architectures.

E.g. constructing and comparing descriptions of possible future actions, needs a "workspace" for items of varying complexity.

Some kinds of neural net require mechanisms supporting continuous variation.

Some kinds of manipulation require an engine able to construct and manipulate "Fregean" structures, with hierarchic function plus arguments decomposition. (E.g. f(g(a, h(b,c)), h(d,e)))

Contrast requirements specified (a) in terms of a virtual machine architecture (b) in terms of physical mechanisms.

A VM SPECIFICATION **might mention a strict stack discipline for procedure activations, with local variables and return address in each stack frame.**
A PHYSICAL SPECIFICATION **might mention fast special purpose registers, etc.**

How much the properties of a particular VM can be decoupled from properties of the physical implementation will vary.

How much of a VM is implemented in the "external" environment will vary. (E.g. pheromone trails used by insects.)

We don't yet know whether there are important types of information processing that computers cannot do — we don't yet know what computers can do.

Varieties of information processing architectures in organisms

Not all organisms can do all the things listed previously.

Everyone knows that organisms can differ in their size, their physiology, their habitats. their behaviours, their social organisation.

Many researchers do comparative studies, and discuss how these things evolved.

Differences in their information processing functions and architectures and how they evolved are not acknowledged to the same extent.

E.g. the chapter on evolution of memory in S.Rose *The making of memory*, 1993, (excellent book) is mainly about evolution of physiological mechanisms and behaviours.

Rose, like many others, seems to think that “information processing” refers only to what computers viewed as bit manipulators do, apparently unaware that even in computers there are many varieties of information processing in different sorts of virtual machines.

Such views could obstruct attempts to study natural information processing architectures and their evolutionary and developmental trajectories.

Testing/checking descriptions of virtual machines in organisms

It’s hard to test a theory about information processing in a system you have not designed.

Hypotheses about VM architecture and information processed are not usually inferrable from descriptions of physical architecture and physical behavioural descriptions. (Compare hypotheses about types of energy in a system.)

E.g. any observed behaviour can be explained by infinitely many different algorithms.

Mappings between VM components and physical components can be indirect and variable.

Given a particular information processing architecture, we cannot foresee all the forms of physical implementation that will be discovered in the future.

Unknown types of implementation may already be used in organisms, “discovered” long ago by evolution.

E.g. we don’t yet know all the functional roles of chemical processes in brains.

DO NOT EXPECT SIMPLE SCIENTIFIC METHODOLOGIES TO BE RELEVANT
E.g. “operationalism”.

Instead use a flexible hypothetico-deductive method, employing a variety of different criteria for comparing and evaluating theories.

Theories can be defeated by better theories, experimental results, internal inconsistencies, etc. but can never be proved.

Evidence in favour will only be circumstantial, never conclusive.

Evolvability could be an important constraining criterion for theories of natural information processing systems.

Do not expect concepts to be precisely defined independently of the theories that use them.

Deep theories will refer to various types of information, types of processing, types of mechanisms, types of uses of information, types of change in information processing, types of implementation.

More rigid philosophical requirements would have ruled out the deepest parts of physics, leaving only shallow studies of correlations.

EVOLUTION OF NATURAL INFORMATION PROCESSING ARCHITECTURES

Work in Computer Science, Software Engineering and Artificial Intelligence has produced many different kinds of VM architectures.

The details are very different from natural architectures, which are generally far more flexible and robust, though usually only within circumscribed domains, though we can relate broad categories to both natural and artificial systems.

AI researchers investigate various kinds of reactive architectures, deliberative architectures, reflective architectures and hybrid types.

***Stateless* reactive architectures are very rigid and of limited use. Architectures where some reactions produce internal state changes influencing future reactions can be very effective for many tasks.**

The vast majority of organisms are "reactive" in that sense: they produce internal or external reactions directly triggered by internal or external states and state changes.

Far fewer organisms have "deliberative" mechanisms, able to construct, compare and select more or less complex descriptions of possible states of affairs or actions (internal or external).

We need to understand the evolutionary trajectories that made possible more and more sophisticated reactive architectures, followed in some cases by more and more complex deliberative architectures.

In principle reactive architectures suffice

Any behaviour observed in any machine or organism over any time period, could, in principle be produced by a reactive system.

However, a reactive system and a deliberative system will differ

(a) in the internal VM processes that occur

(b) in the storage requirements to support the same set of observed (or counter-factual) behaviours

(c) in the evolutionary history or training history required to produce the same range of capabilities.

A biological system that can induce and re-use generalisations in order to cope with novel problems or contexts will not require the same storage capacity or individual history or evolutionary history as one that has only pre-existing solutions selected in a reactive manner.

That is the main type of evidence supporting hypotheses that apes, humans, etc. sometimes use deliberative, not only reactive, mechanisms.

However it is important to note that such evidence is only circumstantial.

What am I proposing?

I am suggesting a possible collection of themes that might be included in an AIBACS initiative. This is not intended to exclude other topics.

The problems outlined here are hard research problems.

They require imaginative cross-disciplinary collaboration from very broad-minded researchers willing to take risks.

There are three closely related core problems:

1 What sorts of evolutionary processes could produce human and other sorts of minds? (e.g. microbe minds, insect minds.)

2 What sorts of minds has evolution actually produced?

3 What other sorts of minds are possible?
(Could they be produced by biological evolution,
and if not why not?)

Why ask about minds, not brains?

The questions are partly about brains, and other devices, but primarily about the information processing virtual machines that are *implemented in* brains. I.e. minds.

We need to distinguish *virtual* machines that process information from the *physical* machines that “implement” those virtual machines.

Reference to virtual machines allows us to formulate deep explanatory principles whose generality is lost if we talk only about physical implementation details. Example:

> Instances of Linux running on a PC, a Sparc, an Alpha, a HPPA, an SGI, a PowerPC, have the same virtual machine components doing the same things (e.g. forking, using pipes, reading files, etc.), despite great physical differences at the CPU level (including older and newer generations of CPUs).

There can also be differences, e.g. due to different cache-sizes, different arithmetic capabilities.

Question: are biological virtual machines (minds) more closely tied to their physical implementations?

It’s an empirical question, hard to study without prejudice. Perhaps biological VMs are partly tied to implementations, but not wholly. Electronic substitutes might be useful sometimes.

How can we study biological virtual machines?

The following are among the approaches often advocated:

1. Bottom-up: Find out what the building blocks (atoms, molecules, blood vessels, neurons, muscles, etc., are capable of, and how they are put together in brains and bodies, then try to model or replicate the mechanisms).

2. Top-down: Start from the various things we know animals can do, and as engineers try to design machines with similar capabilities.

3. Procrastinate: Postpone all design work until we have collected lots more information from psychology, brain science, ethology, etc... (or until philosophers have clarified the nature of virtual machines and how they relate to physical machines).

4. Evolve: Try to use evolutionary computation techniques to replicate what evolution has achieved (at a suitable level of abstraction).

5. Analyse: Attempt to understand the processes of evolution that led to existing architectures as a way of understanding the architectures.

6. Give up: Argue that the whole task is impossible and give up.

7. Synthesize: Combine all the approaches (except 6).

A Recommended Approach

It is unlikely that any of the above approaches can succeed in isolation.

Recommendation: combine them all

(leaving out arm-chair arguments about impossibility, except in the pub).

This does not mean that every researcher must adopt all the approaches.

However:

DIFFERENT RESEARCHERS ADOPTING DIFFERENT APPROACHES SHOULD SHARE INFORMATION, THEORIES AND PROBLEMS, AND COLLABORATE IN ATTEMPTS TO SYNTHESISE.

WARNING: part of the problem is conceptual confusion.

E.g. consider the difficulties and disagreements that arise in explicating our concepts of "intelligence", "learning", "emotion", "belief", "desire", "consciousness", "function", "cause", "computation", "representation", etc.

WHY STUDY THESE TOPICS?
THERE ARE DIFFERENT MOTIVES

Motivations for investigating computational models of mind

(i) **Science:** Studying natural minds as something to be modelled and explained, including explaining how they might have evolved, etc.

(ii) **Engineering:** A desire to produce new kinds of useful machines, with capabilities of humans and other animals.
(NOTE: we already have calculators, spreadsheets, mathematical tools, ... doing things previously done only by humans.)

(iii) **Entertainment:** A desire to produce new kinds of computer-based entertainments where synthetic agents, e.g. software agents or "toy" robots, produce convincing human-like behaviour.
(A multi-billion-dollar fast-growing industry.)

(iv) **Education:** Using models of type (i), (ii), (iii). in educational tools for trainee psychologists, therapists, etc.

(v) **Therapy:** ???

(We cannot hope to produce successful educational strategies or therapies without a deep understanding of the information processing architectures with which we are engaging.)

**The conceptual requirements for these objectives are different.
Also modes of evaluation of research.**

E.g. “believable” behaviour in synthetic characters in computer entertainments could be produced by widely different models, including at one extreme very large, hand-coded lookup tables specifying what to do when.

But in the long run a deep and accurate scientifically motivated model of the first type may be required for effectively achieving the more practical goals including types (ii) and (iii)

For now we address only (i), while keeping an eye on the requirements for the others.

A very simple toy demo of type (iii)/(iv).

(Show simulated sheepdog demo: a purely reactive system:
http://www.cs.bham.ac.uk/research/poplog/sim/teach/sim_sheepdog.p
Alas no time for a demo of a SHRDLU-like deliberative system.)

Studying information processing systems

In previous centuries scientists studied machines that

1. manipulate and transform MATTER (e.g. diggers, ovens)
2. manipulate, transform, store, and use ENERGY (e.g. levers, motors).

During the last century we have begun to understand machines that primarily manipulate INFORMATION, though they do this by being *implemented in* machines that manipulate matter and energy (and currents, voltages, etc. etc.)

A marble rolling down a helter-skelter or water flowing to the sea can be construed as using information about the environment to select between possible motions. However this view (Aristotle's?) adds little to the understanding we get by simply using the laws of physics to predict their motion.

That does not mean the marble is not an information processor: it is just a very simple, rather uninteresting, kind of information processor, as a circle is a limiting case of an ellipse.

What other kinds of information processors are there?

More complex information processing systems

In contrast with the marble, if we described the workings of the following:

- a society,
- a mind,
- a computer operating system, compiler, word processor,

ONLY in terms of the physical implementation, then we would fail to express important features and generalisations about the behaviour of the system.

E.g. ignorance can cause poverty; poverty (or at least relative deprivation – Durkheim) can cause crime; desires and beliefs can cause intentions which can cause actions; discovery of syntactic errors can cause compilation to be aborted.

Treating social systems, humans, software machines, as merely physical systems ignores the fact that, besides using up energy, moving physical objects, switching neural or transistor states, they also acquire, store, transform, and use information about themselves, their environments, the actions available to them, their histories, etc.

This may be an important aspect even of the behaviour of microbes (and not only their reproduction, which involves genetic information).

Nothing magical or mystical is required

An information processing virtual machine (IPVM) does not require some kind of spiritual substance that can survive the destruction of the physical body, as was once thought.

IPVMs exist in virtue of the organisation of physical systems that implement them.

They can use information about the current state, about previous states, about nearby phenomena, about remote phenomena, about actual events, about possible events.

Compare:

• A time-sharing operating system's scheduler interrupts processes partly on the basis of records of *previous* machine usage.

• An optimising compiler can transform a partially compiled program (intermediate data-structures in the compiler's virtual machine) on the basis of information it has derived about the structure of the program or on the basis of analysis of what *would happen* under certain conditions that could arise while the program runs.

Biological systems process information in far more varied, complex and subtle ways: mostly not yet understood.

Compare: Norbert Wiener *Cybernetics: or Control and Communication in the Animal and the Machine* 1948, 2nd ed 1961.

MANY MINDS ARE MORE COMPLEX THAN OPERATING SYSTEMS

Percepts, beliefs, desires, pains, plans, etc, are components of information processing virtual machines so far found only in biological systems. (Simplified versions exist in some robots.)

Different kinds of IPVMs exist in different sorts of animals.

E.g. W. Kohler (*The mentality of Apes* 1927) discovered what appeared to be creative problem solving capabilities in some chimpanzees. In *The Nature of Explanation*, 1943, K. Craik discussed reasons for the biological usefulness of such deliberative capabilities. Also Popper.

Only in the last 40 years have we begun to understand some of the complex architectural requirements for problem solving virtual machines e.g. the need for different sorts of long term and short term information stores, for various kinds of information manipulation, and for means of avoiding or reducing search.

Very few animal species have such capabilities. Why?

We should try to understand precisely what the differences between different animal information processing architectures are, how they evolved, how they develop in individuals, colonies or societies, and what sorts of mental processes they do and do not support.

Some methodological points

The task cannot be done by direct observation.

You can't simply *observe* whether genes, microbes, insects, humans or ecosystems, acquire, store, transform, transmit, and use information, any more than you can observe the changes in chemical bonds as water forms from hydrogen and oxygen.

Objection: If so many diverse things are information processors doesn't the idea become vacuous?

No more than the idea of energy, which is everywhere.

That leaves the task of investigating and analysing the variety of information processing systems, distinguishing:
- **The types of things they can do**
- **The types of contexts in which they can do them**
- **The types of ways they can adapt, develop, learn, evolve**
- **The types of architectures and mechanisms (including representations algorithms, social organisations, physical interfaces, etc., etc.) which make those capabilities possible.**

If biological evolution produced different types of IPVMs at different times, including reactive, deliberative and other architectures, we need to investigate how and why — and why many different kinds flourished simultaneously, instead of one kind dominating. This requires analysis of trade-offs.

WARNING:
When looking at particular cases,
Don’t ask:
Is this or isn’t it information processing?
Instead ask:
Which kind of information processing system is this?

It may be a pretty trivial kind, e.g. in a thermostat, or a very sophisticated kind, e.g. when human visual systems enable recognition of mood or emotion in others.

But differences in *degree* (e.g. of complexity) are not what we are looking for, for then we miss the more important differences in *kind* (e.g. kinds of capabilities, adaptations, representations, algorithms, architectures, ...)

Dichotomies are often useless, whereas taxonomies drive deeper understanding.

A GOOD THEORY MUST ALLOW FOR VARIATION

Many theories (e.g. of emotions) implicitly assume that there is only one kind of architecture.

But natural information processing systems, whether perceptual systems, learning systems, action control systems, motivational systems are very diverse. We need to understand this diversity.

There are different sorts of variation:

- **Across species,**
- **Within species,**
- **Within an individual during normal development**
- **After brain damage**
- **Across the natural/artificial divide.**
- **Perhaps across planets (grieving, infatuated, curious, Martians?)**

Note that these are not variations in *degree* but variations in *kind*: involving different kinds of capabilities, different kinds of information, different kinds of architectures.

WHY STUDY VARIETIES OF ARCHITECTURES?

Philosophers want to understand the space of possibilities, and their implications.

Engineers look for both re-usable old solutions and novel ones.

Scientists are likely to miss things if their search is too focused.

You don’t really understand an architecture unless you understand its advantages and disadvantages compared with neighbours in design space. (In different contexts, etc.)

In any case there are very different cases in nature, even among humans, e.g. infants, children, altzheimers patients, etc.

Some confusions to be avoided:
Components of information processing architectures need not be physically identifiable (certainly not easily identifiable)

When Newell and Simon talked about the need to study *physical symbol systems* they caused considerable confusion: they should have called them *physically implemented* symbol systems.

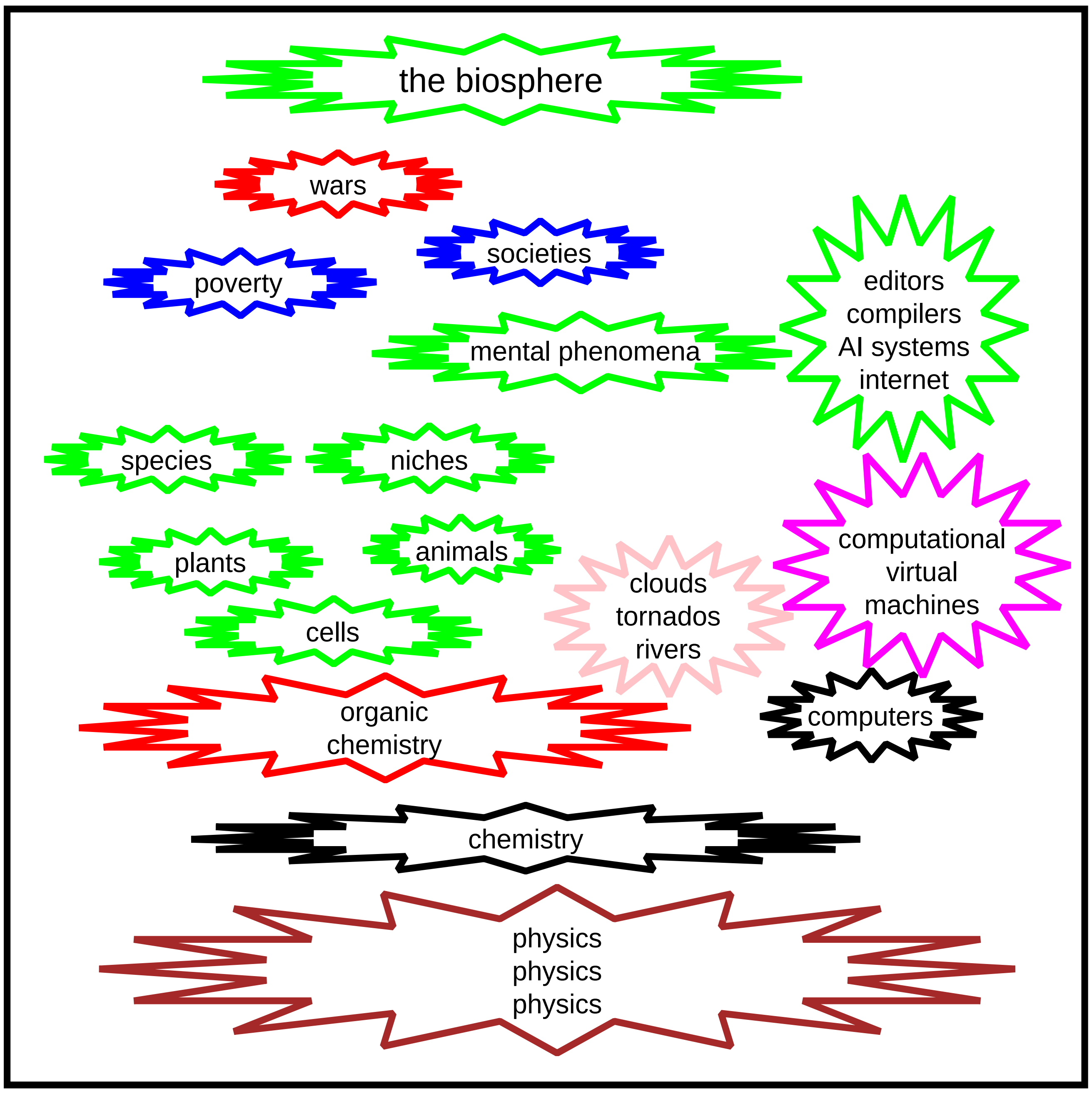
But the information encoded in a trained neural net may not be best viewed as a “symbol system” in either sense. (It depends on the neural net.)

The net is still an information processing mechanism, doing one kind of processing during training (e.g. finding abstractions) and another kind during use (e.g. classifying, controlling, interpolating, extrapolating, etc.).

There are many ways of encoding, transforming, storing, transmitting, and using information, using many forms of representation, and many types of structure-manipulating mechanisms.

Avoid assuming that any currently well understood subset (e.g. various types of neural nets, condition-action rule systems, computer data-structures, systems manipulating logical formulas, natural language systems, image arrays, continuous dynamical systems) exhaust the possibilities.

WE BARELY UNDERSTAND THE VARIETY OF VIRTUAL MACHINE ARCHITECTURES



Emergent virtual machines are everywhere

How many levels of physics will there be in 500 years time?

Many kinds of causally efficacious “emergent” virtual machines, operating at different scales and different levels of abstraction.

- Is there a well-defined “bottom level” ? We don’t know whether physics has a well defined bottom level. It seems to have several levels.
- Causality operates at all levels. (Poverty can cause crime)
- Causality can be “circular”.

  (E.g. physical processes in a computer cause events in a computational virtual machine, and non-physical events in the virtual machine, e.g. a queen capturing a bishop, a theorem prover finding a contradiction, cause physical events, in the computer, and on the screen.)
- Levels need not form a rigid hierarchy.
- The biosphere is a richly interacting collection of myriad physical and virtual machines running concurrently.

Recommendation: for now, just treat all those as facts to be explained by deep scientific theories. Leave philosophical arguments about their interpretation to trained philosophers.

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.

COULD EVOLUTION BE A PURPOSEFUL WATCHMAKER?

TRADITIONAL DARWINIAN VIEW:

Evolution is a "blind" process using only random processes which may or may not produce good solutions to biological problems.

This is a "local hill-climbing mechanism".

Now consider a search space with 1000 binary steps, and a thousand million options considered every microsecond.

Exhaustive search requires this number of years:

$(2^{1000})/(1000000000*1000000*60*60*24*365)$,

which is:

3397731504268982000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000

years!

Estimated age of the earth: 4500,000,000 years, approximately.

Question:

Could mere local hill climbing search such a space as biological evolution does?

Or might there be higher level information processing mechanisms *implemented* within the Darwinian mechanisms that control and guide the search?

Maybe we need to understand different types of trajectories in the space defined by Darwinian mechanisms, and how they are controlled.

(This is not anti-Darwinism.)

MORE QUESTIONS

How many design decisions are there in a typical organism, e.g. a flea or a human?

What is the search space like?

What proportion of the space has to be explored ?

Does it have a structure that facilitates the search for good solutions?

How? Is co-evolution part of the answer?

Mutually interacting trajectories vs mere parallelism.

Could the intelligence of the organisms that are produced by evolution be part of the answer?

E.g. mate selection, breeding of plants and animals, and now genetic engineering.

CONJECTURE:

There is some higher level architecture at work which supports something more “purposive”.

It doesn’t involve conventional intelligence, goals, plans, etc.

Perhaps a kind of feedback mechanism: e.g. more like an amplifier.

It may be either

• a contingent consequence of the details of evolution on earth,

or

• a necessary (mathematically demonstrable) consequence of some of the properties of Blueall evolutionary processes in some interesting class of processes.

This is not a new idea:

Compare ‘Laws of form’ (Kauffman, Goodwin).

HOW COULD MULTI-LEVEL DARWINIAN EVOLUTION OPERATE? PERHAPS DARWINIAN EVOLUTION USES HIGH LEVEL STRUCTURES AND PROCESSES

POSSIBLE MECHANISMS?

- Feedback loops between niches and designs in co-evolution
- Mate selection
- Social mechanisms (Spartans killing weak offspring)
- Advances in knowledge leading to genetic engineering.
- A hidden global architecture in the total evolutionary system, permitting a mixture of reactive and deliberative mechanisms (e.g. an ecosystem experimenting to solve a problem?)

IS THIS TOTALLY CRAZY ????

Two forms of answer:

- **Experimental simulations showing the possibility of the mechanisms**
- **Mathematical analysis showing the inevitability of the mechanisms (Based on a new conceptual framework)**

Either of these ultimately needs to be related to empirical data.

TOWARDS A CONCEPTUAL FRAMEWORK: A FIRST PASS

- A *design* specifies a class of architectures
  (each of which can have many subclasses and instances)
- A *niche* specifies a class of environments and criteria for evaluating/comparing designs.
  (Engineers talk about a “requirements specification”)
- Every combination of design instance and niche determines an evaluation.
  (However, this need not be a simple numeric fitness function)

When the niche changes (e.g. because another design has changed, or the physical environment has changed) this can alter the evaluation, i.e. the quality of the design for the niche.

Likewise when the design changes, the evaluation can change.

This can produce “pressure” to change the design.

That alters the niches for other designs, and their evaluations,

.... and so on.

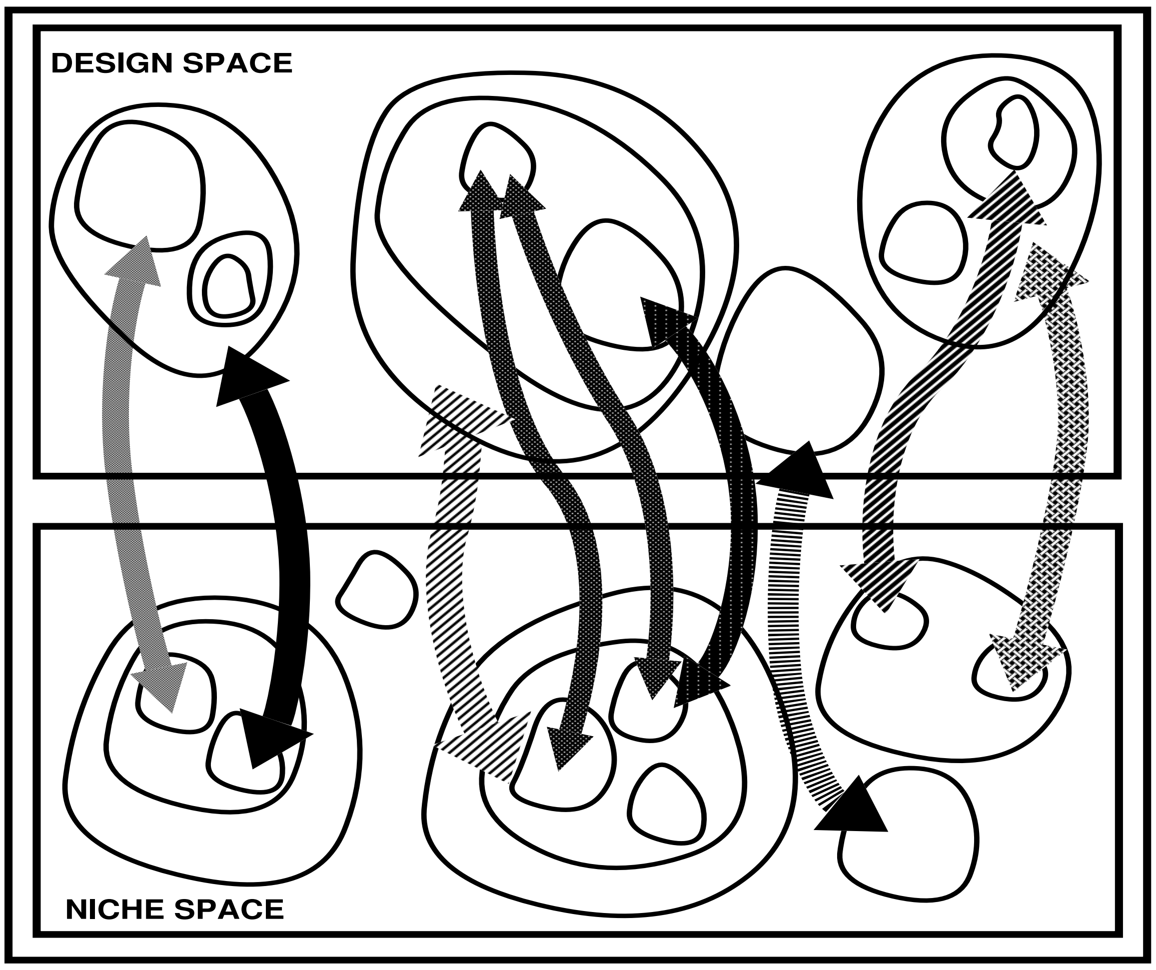
I.e. there are multiple interacting feedback loops within different subsystems. (Otherwise known as co-evolution.)

We need good ways of thinking of possible trajectories in niche space and design space

Perhaps we can then begin to understand the dynamics of evolution, at an appropriate level of abstraction.

We may need new kinds of mathematics, e.g. to characterise the structures of the spaces and paths within them.

Design space, niche space and evaluation relations



Relations between designs and niches are complex and varied.

Instead of 'fitness'* VALUES *we need fitness* DESCRIPTIONS*.

Different sorts of arrows represent different types of fitness relations.

One sort of description might be a vector of values (or labels), as in Consumer Association reports.

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: i.e. different generations.

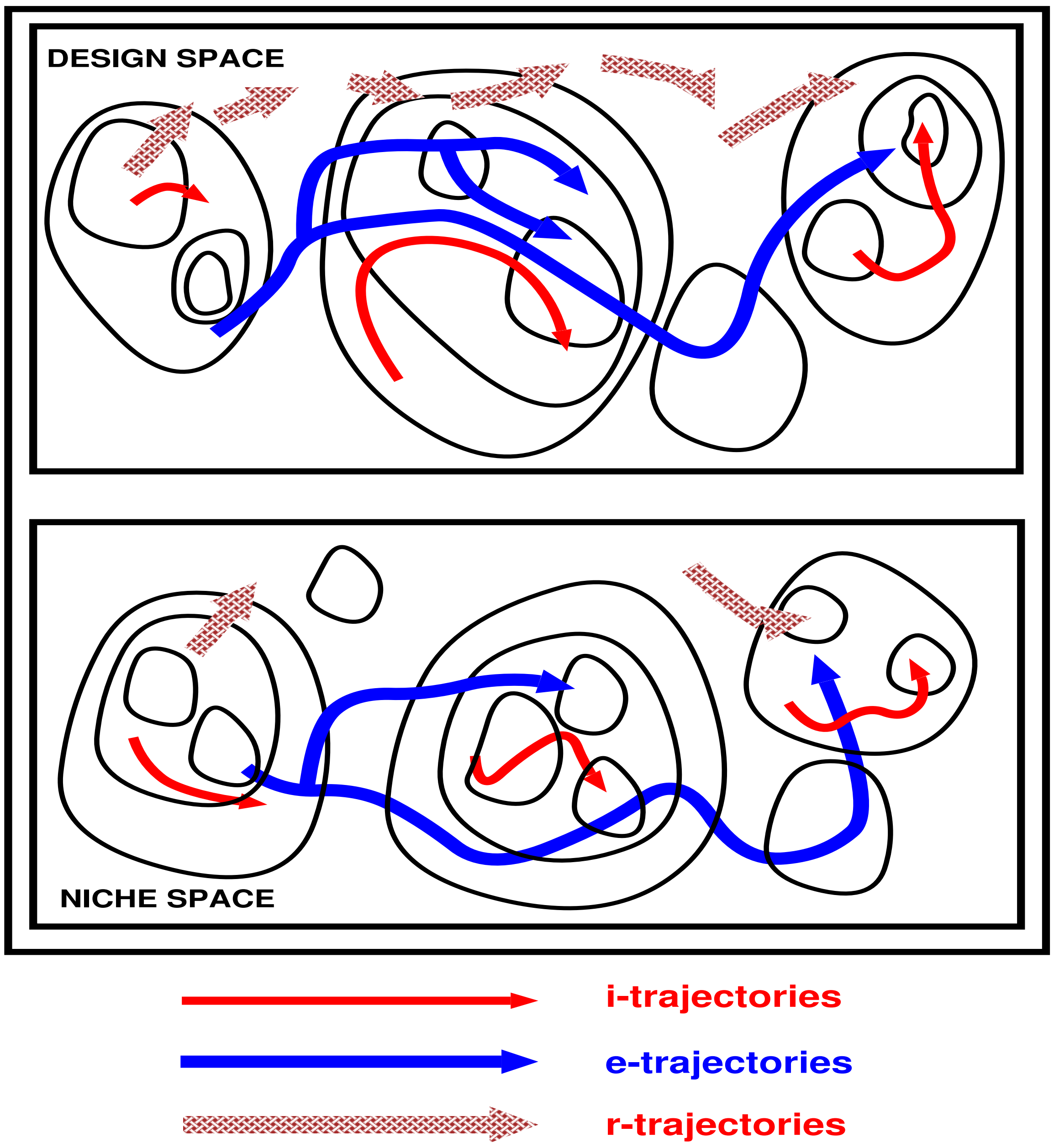
r-trajectory: possible for a system being repaired or built by an external designer whose actions turn non-functioning part-built systems into functioning wholes.

c-trajectory: evolutionary trajectory making use of cognitive capacities of individuals (e.g. mate selection).

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

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

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



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

But don't forget that biological evolution is discontinuous.

NOTES ON TRAJECTORIES

Note 1: what is an e-trajectory or s-trajectory from one viewpoint may be an i-trajectory from another: viewing a species or a society as an individual.

Note 2: Trajectories occur both in niche space and in design space.

Note 3: Understanding the dynamics of ecosystems requires understanding the feedback loops between different trajectories at different levels of abstraction in both spaces.

Note 4: We can also construe a complex organisms as a sort of ecosystem, in which parts, or rather their designs, co-evolve.

The last point leads to a new view of the architecture of an individual. E.g. perceptual systems may have parallel sub-systems concurrently serving the needs of concurrently active internal sub-systems.

More on this later.

Biological evolution:
Multiple interacting e-trajectories,
later using i-trajectories,
then c-trajectories,
and s-trajectories,
and now also r-trajectories

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?

Must organisms with deliberative capabilities be at the top of a food pyramid?

Compare producing large numbers of cheap, expendable, “stupid” offspring.

(Using “stupid” as an abbreviation for “lacking various kinds of learning, predictive, planning, self-monitoring, etc. capabilities”.)

TOWARDS A TOPOLOGY FOR DESIGN SPACE AND NICHE SPACE

Both spaces are mathematically very complex

- Both may have some continuous regions and also many discontinuities.

  (Remember Darwinian evolution is inherently discontinuous)

- Regions in both spaces are defined by abstract specifications or collections of specifications.

- The sizes of discontinuities vary according to level of abstraction.

  (Two more legs, vs slightly longer legs)

- Trajectories in both spaces can be defined at different levels of abstraction.

- In many cases the specification of a niche depends in part on the specification of a design. E.g. food-getting requirements depend on existing food-digesting mechanisms

- This leads to the view of an organism as itself some kind of ecosystem, composed of co-evolved sub-organisms.

ASPECTS OF THE LOGIC OF ‘BETTER’
THE IMPORTANCE OF TRADE-OFFS

In general there will not be an answer to the question: does this change (e.g. a mutation) lead to higher fitness?

Rather there will be trade-offs: the new variant is better in some respects and worse in others, and for a given respect the order may depend on circumstances.

This should not be confused with a “neutral” change which makes no relevant difference to the individual’s abilities to fit the niche.

Sometimes there is a partial ordering of the fitness descriptions, and sometimes not even that, because there is no way to combine the different dimensions of comparison.

Transitivity violations can occur:
Design A might be better than B in one respect, B better than C in another and C better than A in a third.

Likewise in different contexts.

IMPLICATIONS OF TRADEOFFS

Tradeoffs between dimensions in fitness vectors may be exploited by the formation of cooperative behaviours and division of labour.

Two individuals that excel in different ways (e.g. hunting and farming) may together be more competent than two with equal but intermediate levels of expertise.

This is particularly true of individuals in a social group requiring many kinds of expertise.

Useful division of labour can also occur across species.

There may be “symbiotic” genes for collaborative sub-organisms and for collaborative complete organisms.

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
- A chimpanzee "conscious" of being threatened by another
- 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...)**

A trade-off between competence at birth and adult competence.

One cost of the latter is nurturing of offspring.

One benefit is acquiring competence that could not have evolved in the history of *THAT* species.

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 (partly) different architecture from a dancer’s.

E.g. the mathematician includes a collection of mechanisms (in the virtual machine architecture) for manipulating formulas, proofs, and other abstract structures.

The dancer may include mechanisms for controlling very complex types of motions requiring specialised coordination between different body-parts and perceptual states.

Both are produced by years of training.

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 mental concepts.

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 (STATELESS OR STATEFUL)
- DELIBERATIVE
- REFLECTIVE (SELF-MONITORING, SELF-CONTROLLING) ...

We can also divide the functionality (three towers):

- SENSORY/PERCEPTUAL SYSTEMS
- INTERNAL PROCESSING
- MOTOR SYSTEMS

These form only a very crude first set of divisions. All of them need further analysis and refinement.

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.**

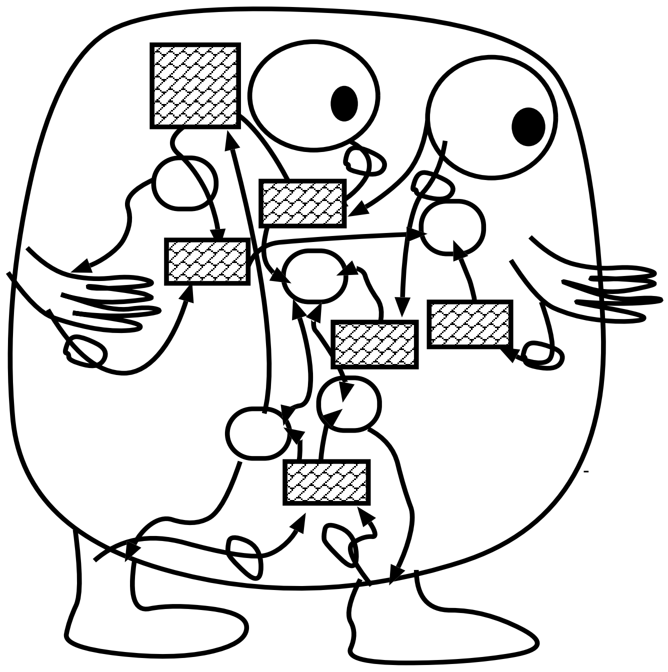
Obvious constraints:

- **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".

WHAT SORT OF ARCHITECTURE?
COULD IT BE AN UNINTELLIGIBLE MESS?



YES, IN PRINCIPLE.

BUT

it can be argued that evolution could not have produced a totally non-modular yet highly functional brain.

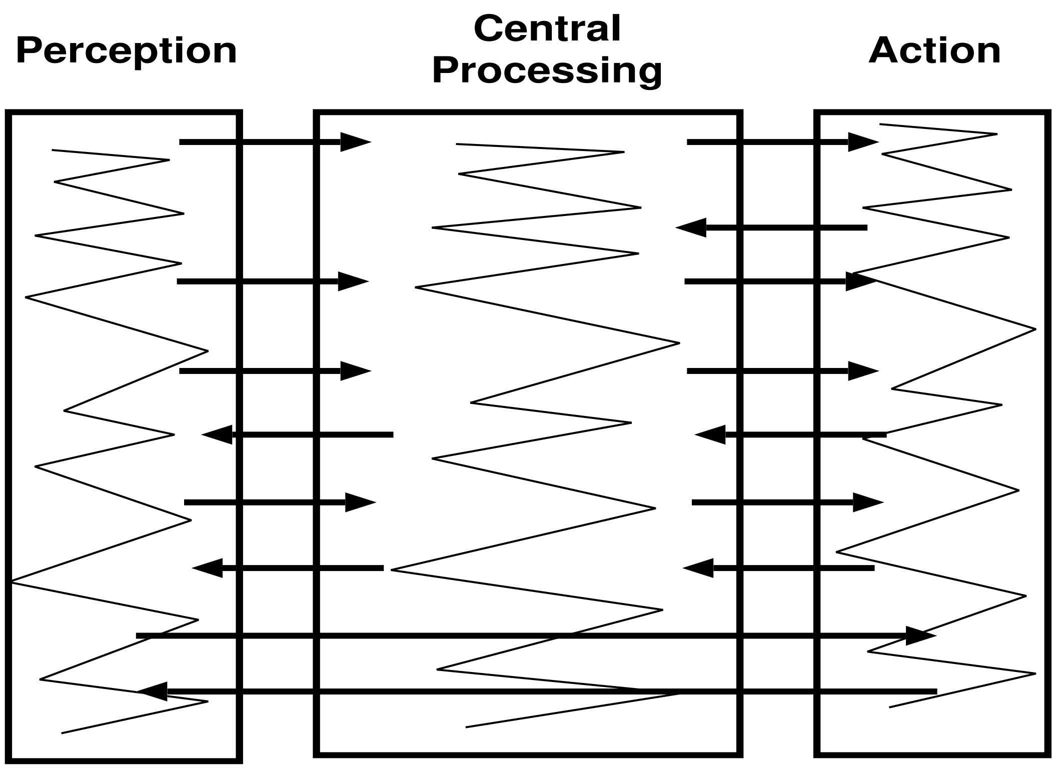
Problem 1: time required and variety of contexts required for a suitably general design to evolve.

Problem 2: storage space required to encode all possibly relevant behaviours if there's no "run-time synthesis" module.

Problem 3: fragility of non-modular system across changes.
(Problem grow worse as complexity of function grows.)

Towards a unifying theory of architectures: For (some) natural and artificial agents

1. THE "TRIPLE TOWER" PERSPECTIVE



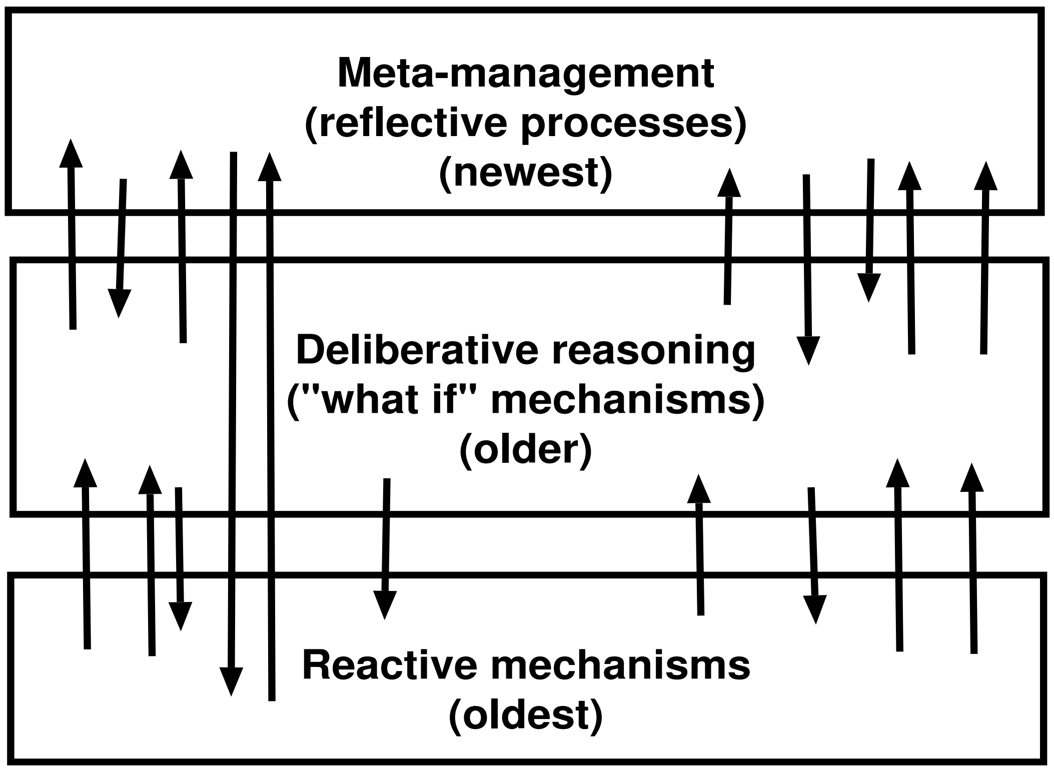
(many variants)
(Nilsson, Albus)

Systems can be "nearly decomposable". Boundaries can change with learning and development. Components can be shared across boundaries.

(E.g. part of visual system sometimes used for reasoning.)

ANOTHER COMMON ARCHITECTURAL PARTITION (functional, evolutionary)

2. THE "TRIPLE LAYER" PERSPECTIVE

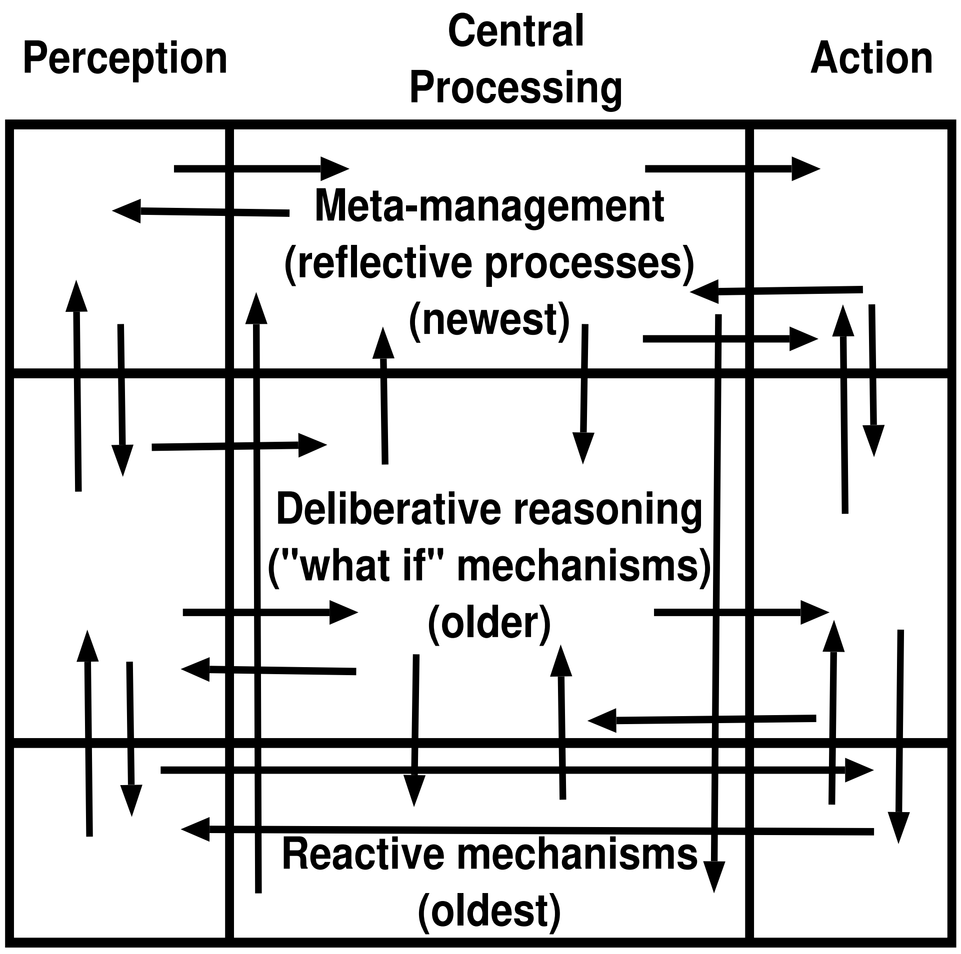


(many variants – for each layer)

Reactive systems can be highly parallel, very fast, and may use analog circuits, with or without internal state changes.

Deliberative mechanisms are inherently slow, serial, knowledge-based, resource limited, with very "expensive" associative memories.

COMBINING THE VIEWS:
LAYERS + PILLARS = GRID



A grid of co-evolving sub-organisms, each contributing to the niches of the others.

An architectural "schema" not an architecture.

Not all instances of the schema will have all components.

E.g. insects probably include just the reactive layer (including various kinds of learning, adaptation, and stored sequences).

The CogAff schema defines a variety of components, and possible information linkages, which may or may not be present in different instances.

It does NOT specify control flow, or dominance of control

Many options are left open.

CogAff subsumes many types of more specific architectures, including “Omega” architectures, subsumption architectures, and H-Cogaff (all sketched below).

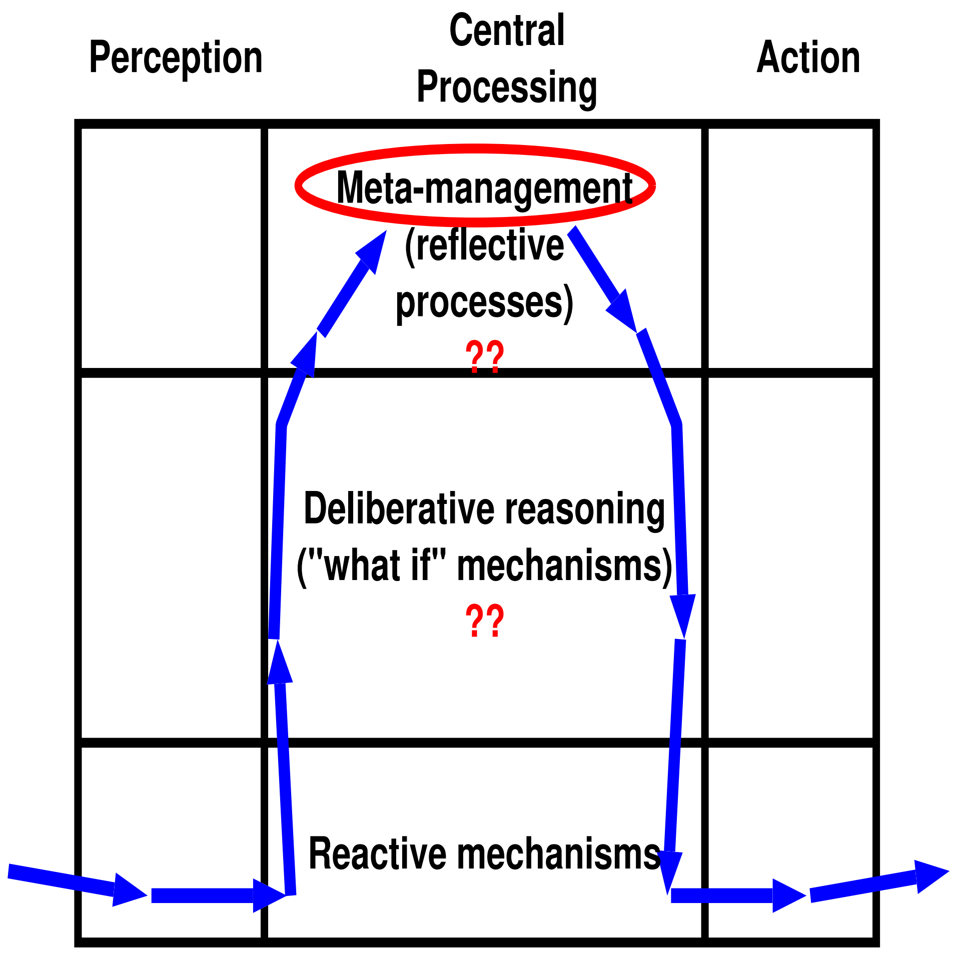
Layered architectures have many variants

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
  – on next slide)

The “Omega” model of information flow



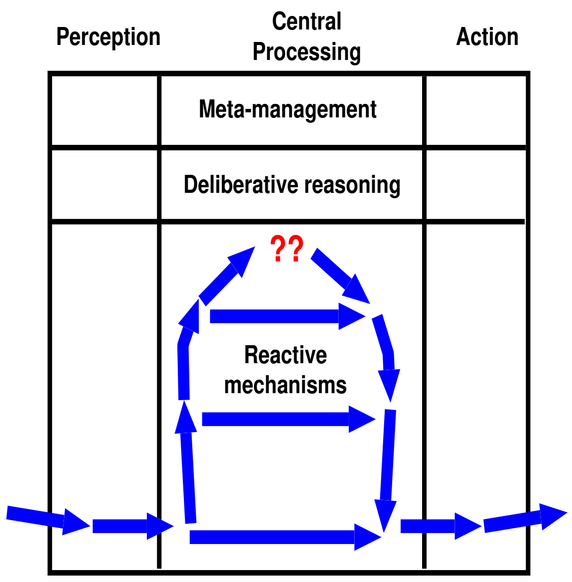
Rejects layered concurrent perceptual and action towers separate from central tower.

I.e. uses only some components and links allowed in CogAff.

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

Some authors propose a “will” at the top of the omega. (Albus 1981)

Another variant (Brooks): Subsumption architectures, with several reactive layers in a control hierarchy



Brooks denies that animals (even humans) use deliberative mechanisms.
(How does he get to overseas conferences?)

Personal opinion: "Nouvelle AI" is of use only in very restricted contexts

The architectures/mechanisms proposed are too obviously incapable of meeting most requirements.

Conjecture: the people who say no planning, deliberation, symbol-manipulation is needed, actually use those methods in managing their own lives.

Likewise defenders of "dynamical systems".

In a sense everything is a dynamical system, though not necessarily usefully described by systems of partial differential equations linking continuous variables. (See debate in *Behavioural and Brain Sciences Journal* Vol 21, No 8, 1998.

What goes on when you think about algebra, or try to find a bug in a computer program, or read a poem?

Shakespeare thought that we are information processing engines:

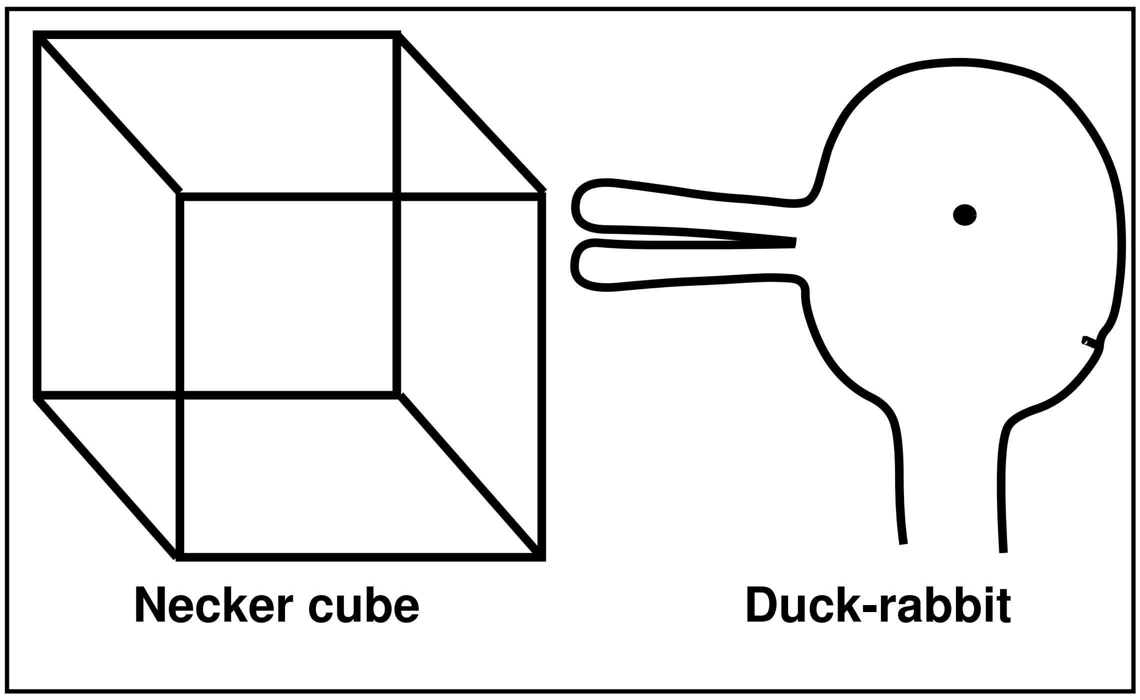
> Love is not love
> which alters when it alteration finds

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.

Levels of abstraction in perceptual mechanisms



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. (Contrast 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...

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 INTERNAL LANGUAGES OF SOME SORT

Talk of *two* visual pathways (ventral, dorsal) looks like a gross oversimplification from this viewpoint: there are many visual pathways serving different sorts of needs. Maybe there are major sub-divisions corresponding to different levels in the CogAff schema.

An extension of Gibson's theory:

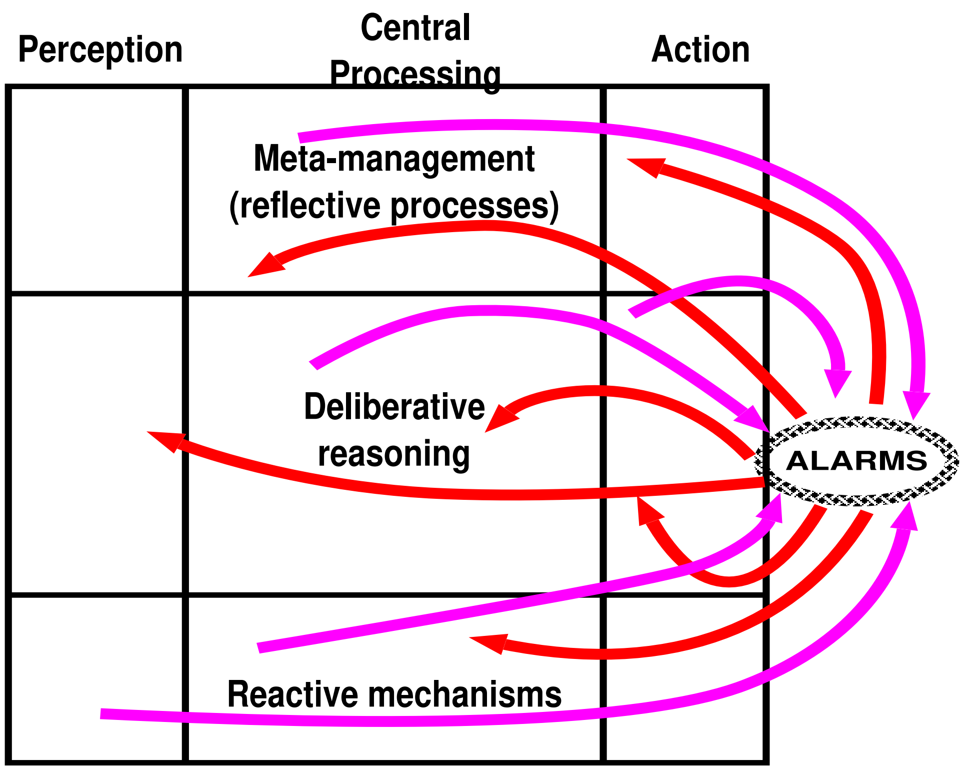
Different sub-systems use different affordances, and different ontologies. (Evidence from brain damage.)

They rely on processing by different virtual machines.

Steps towards an "ecology of mind"

As processing grows more sophisticated, so it can be come slower, to the point of danger.

REMEDY: FAST, POWERFUL,
“GLOBAL ALARM SYSTEMS”



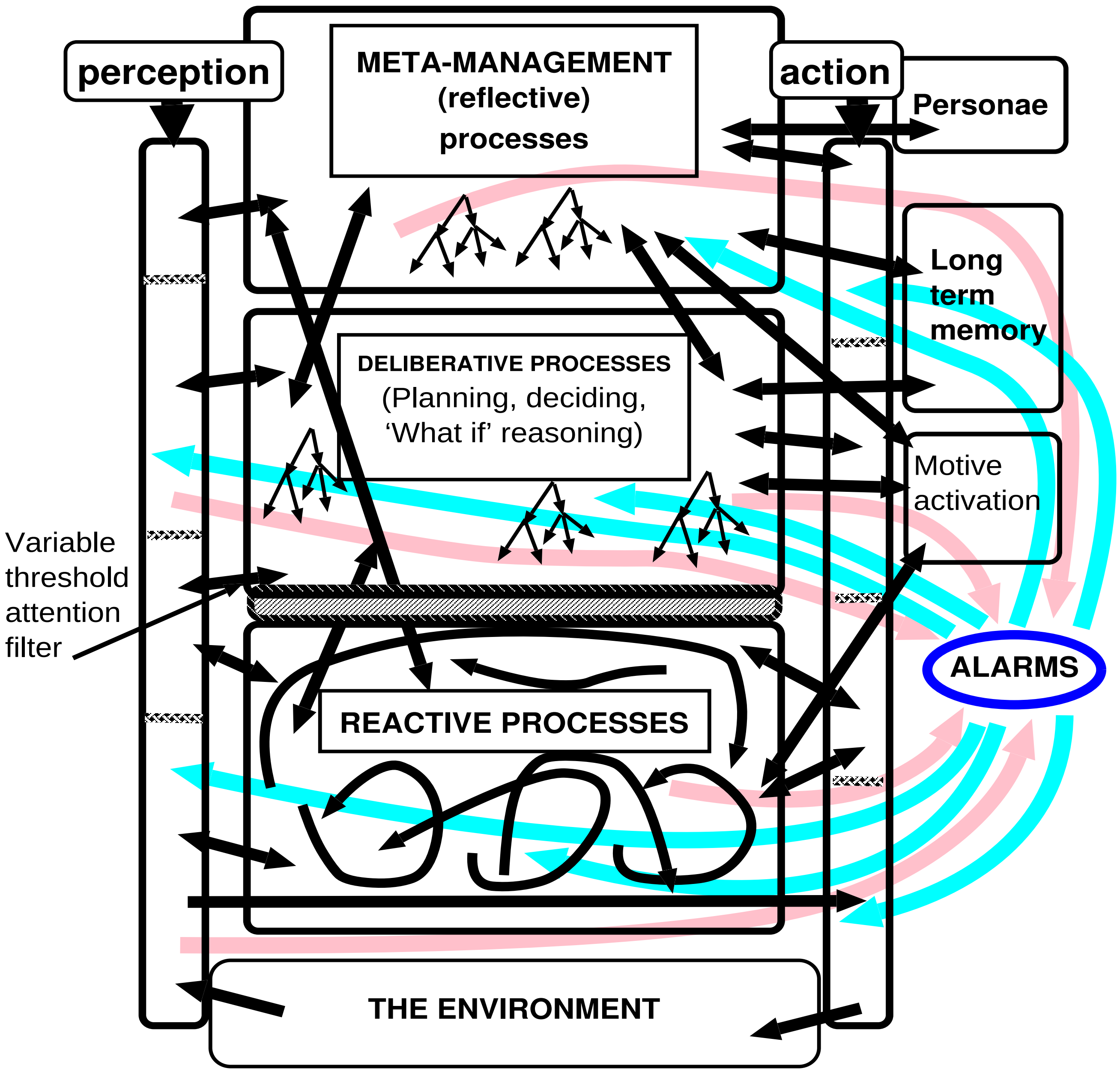
ALARM MECHANISMS MUST USE FAST PATTERN-RECOGNITION AND WILL THEREFORE INEVITABLY BE STUPID, AND CAPABLE OF ERROR!

Note: An alarm mechanism is just part of the reactive layer – drawn separately to aid presentation.

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

E.g. one alarm system or several? (Brain stem, limbic system, ...???)

H-COGAFF: A human-like architecture.
(An instance of the CogAff schema
using all the components.)



Described in more detail in papers in the Cogaff directory:
http://www.cs.bham.ac.uk/research/cogaff/

ONE OR MORE ALARM MECHANISMS (Brain stem, limbic system, blinking reflexes, ...???)

Allows rapid redirection of the whole system or specific parts of the system required for a particular task (e.g. blinking to protect eyes.)

Can include specialised learnt responses: switching modes of thinking after noticing a potential problem.

E.g. doing mathematics, you suddenly notice a new opportunity and switch direction. Maybe this uses an evolved version of a very old alarm mechanism.

The need for (POSSIBLY RAPID) pattern-directed re-direction by meta-management is often confused with the need for emotions e.g. by Damasio, et. al.

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.

(See papers in the Cogaff directory.)

Tertiary emotions (Called "perturbances" in older Cogaff project papers.)

Involve interruption and diversion of thought processes.

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

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.

An organism that does not have meta-management cannot control attention, etc. and therefore cannot LOSE that sort of control, and therefore cannot have tertiary emotions.

It does NOT follow that tertiary emotions are required for intelligent control.

(Damasio's non-sequitur.)

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.

Back to the key idea
THEORIES OF MIND
PRODUCED BY EVOLUTION

An organism with some partial ability to monitor, describe, evaluate its own internal states needs an ontology for that purpose.

The ontology need not be deep or accurate, as long as it works (compare the "naive physics" used by animals to interact with the physical environment).

The same ontology might be capable of being used also to perceive, think about and interact with OTHER **intelligent organisms.**

As scientists developing theories of the architecture of animal minds, we are therefore catching up with, extending, and refining something produced previously by evolution, namely: our own pre-existing implicit, simplified, theories.

Watch this space.