Talk at Newcastle International Seminar
5-8 Sept 2000

Intelligent Beasts and Artefacts
or
How to Turn Philosophers of Mind into Engineers

AARON SLOMAN

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

**Ideas developed
in collaboration with
Steve Allen, Luc Beaudoin, Darryl Davis,
Catriona Kennedy, Brian Logan,
Matthias Scheutz, Ian Wright,
and others in the**

COGNITION AND AFFECT PROJECT

SCHOOL OF COMPUTER SCIENCE
THE UNIVERSITY OF BIRMINGHAM

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

I HAVE ALSO LEARNT FROM MANY OTHERS.

BACKGROUND

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.

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

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

NB: "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".

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

Such a study is facilitated by thinking of overt human language as a *special* case of a more general notion of "language":

> **A (possibly extendable) collection of information-bearing structures embedded in a manipulable medium.**
>
> **Where languages may have many uses that have nothing to do with interpersonal communication: e.g.**

THINKING, REMEMBERING, WONDERING, DESIRING, DELIBERATING, DECIDING, LEARNING, PLANNING, EXPLAINING

To explain all this we need to think about architectures.

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

E.g. because we need to know how to put things together

- **We now have a small set of ideas about architectures (including the Birmingham 'Cogaff' architecture schema)**
- **And a variety of tools for investigating and using them (Soar, Cogent, PRS, ACT-RPM, Sim_agent .....)**
- **But the space of architectures is huge and ill defined**
  - **organised in various kinds of sub-architectures**
  - REACTIVE
  - DELIBERATIVE
  - HYBRID ...

With a very large variety of possible components,

Including:

- PARSERS
- INFERENCE ENGINES
- CONSTRAINT MANIPULATORS
- RULE INDUCTION ENGINES
- IMAGE ANALYSERS
- NEURAL NETS
- EVOLUTIONARY MECHANISMS
- ANALOG DEVICES
- DYNAMICAL SYSTEMS
- MANY KINDS OF TRANSDUCERS (SENSORS AND MOTORS)
- CHEMICAL COMPUTERS

Show examples

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**

More subtle constraints: "what is evolvable".

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

Deep understanding will not come from studying ONE case – a typical adult human mind!

We need to explore alternatives, understand trade-offs.

Let's look at neighbourhoods

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

and learn from their similarities and differences.

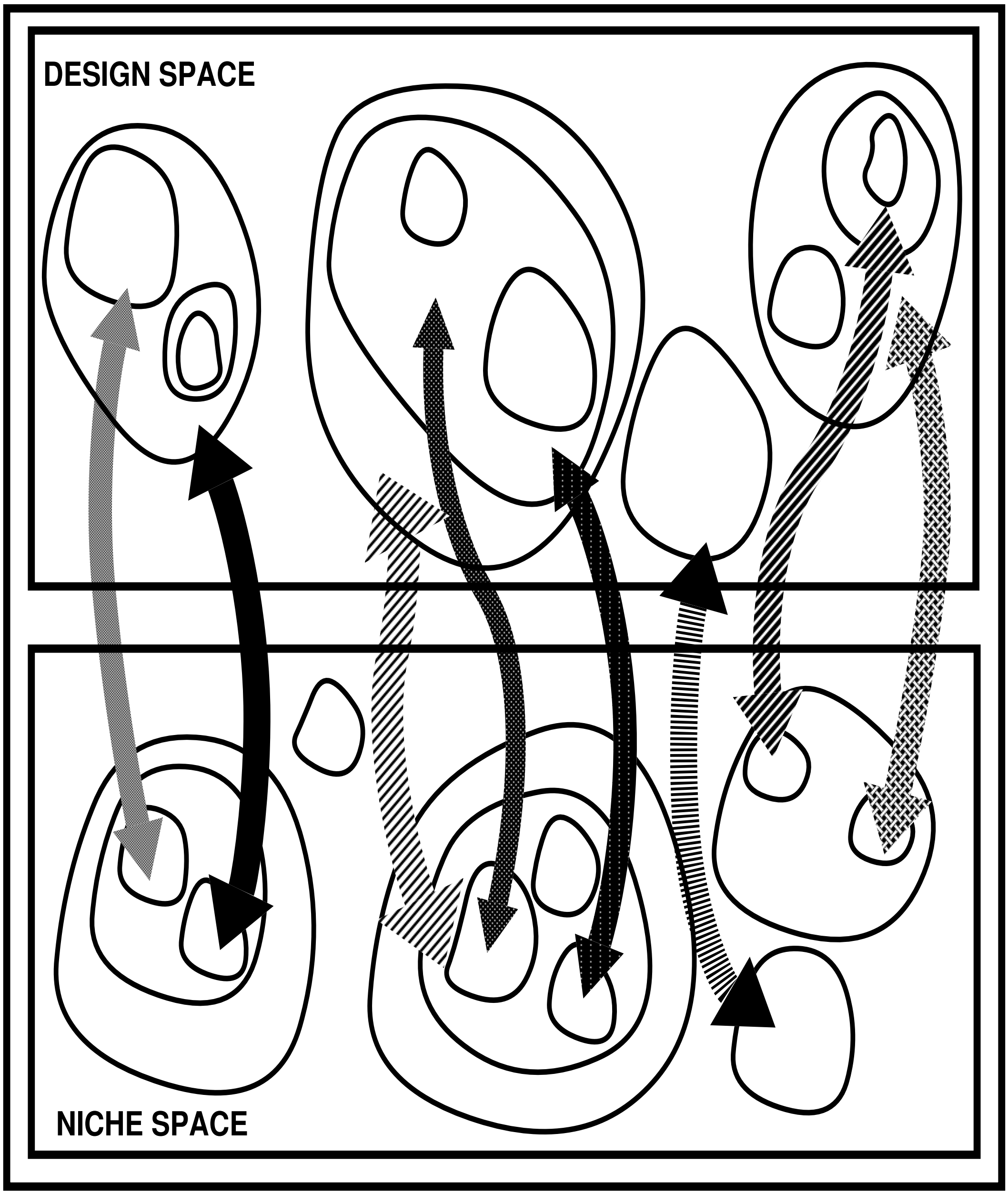
We need to understand different types of *trajectories* through these spaces, in evolution, in individual development, in learning, in cultural change, ...

We need to understand the interactions between the trajectories, i.e. *the many feedback loops* in co-evolution.

We need to understanding 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.



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

We need to understand the structure of both design space and niche-space

This includes:

Analysing the variety of mappings between them (NOT numerical-valued ‘fitness functions’ – perhaps vector valued?)

Exploring the variety of types of evolutionary pressures, and the feedback loops in niche space and design space involved in co-evolution.

Finding out what sorts of evolutionary and developmental trajectories in design space and niche space are possible and how.

(Remember: Biological evolution is DISCONTINUOUS)

Many of our concepts are DEEPLY confused

- **E.g. ‘emotion’, ‘representation’, ‘computation’**
- **We can use architecture-based concepts to refine and extend them.**
- **Compare physics: the architecture of matter**
- **But beware: there’s not just one architecture for mind We need COMPARATIVE investigations**

I.e. collect examples of many types of real phenomena.

Try to build a theory which explains them all!
Subject to constraints from neuroscience, psychology, biological evolution, feasibility, tractability, etc.

ALLOW FOR VARIATION:

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

ANYONE WHO COMES UP WITH ONE ARCHITECTURE FOR MINDS HAS PROBABLY GOT IT WRONG!

ARCHITECTURE-BASED CONCEPTS OF MIND

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

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

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

- **different kinds of moods, motivations, attitudes**

COMPARE THE ARCHITECTURE OF MATTER

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

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

EXAMPLE
WHAT KIND OF MACHINE
CAN HAVE EMOTIONS?

PROBLEM:

MANY different definitions of "emotion".

in psychology, philosophy, neuroscience . . .

with many variants within each discipline

DIAGNOSIS:

Different theorists concentrate on different phenomena.

We need a theory that encompasses all of them.

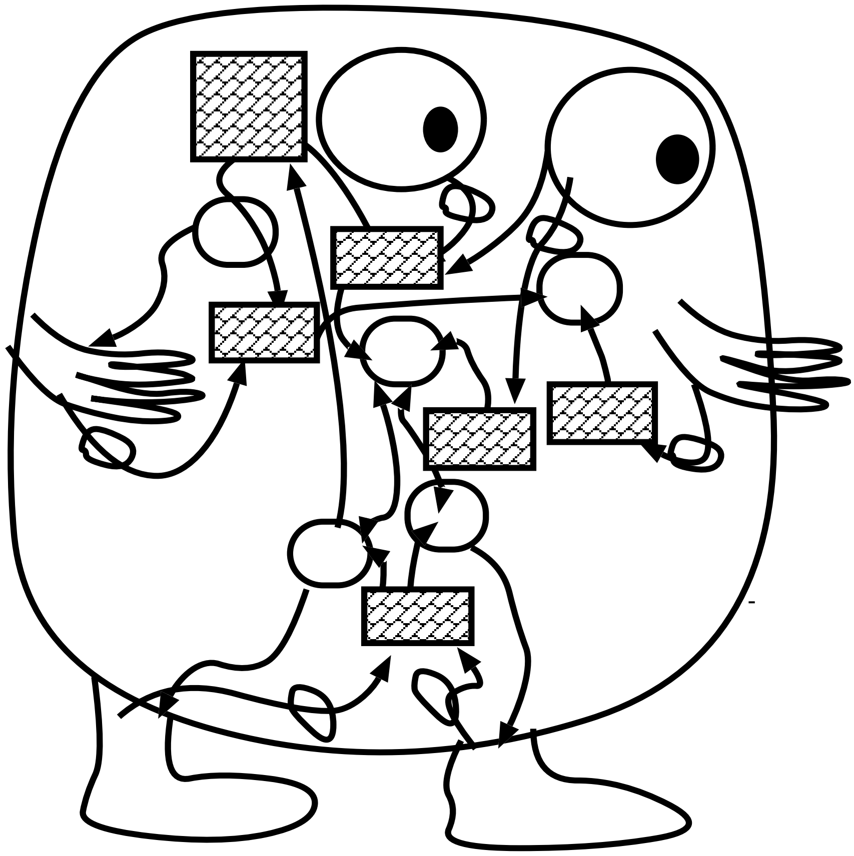
REPHRASE:

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

(DEFINITIONS CAN COME AFTER WE HAVE GOOD THEORIES.)

WHAT SORT OF ARCHITECTURE
CAN ACCOUNT FOR
SUCH PHENOMENA?

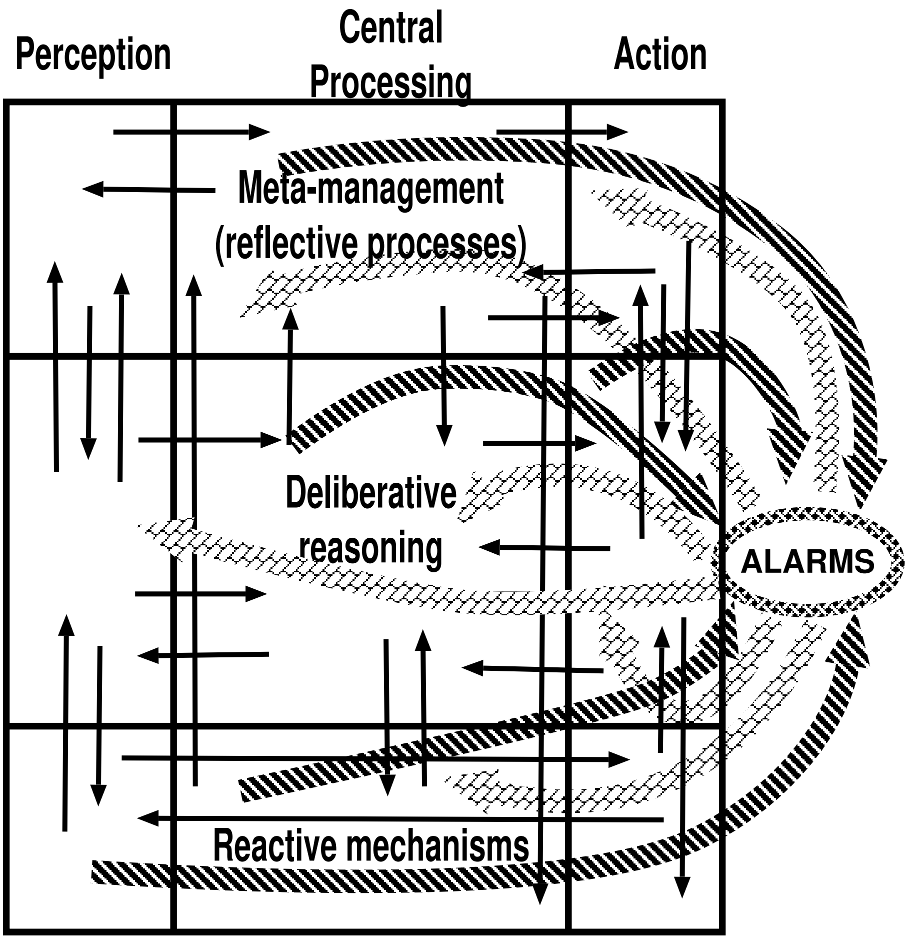
COULD IT BE AN UNINTELLIGIBLE MESS?



Yes, in principle.

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

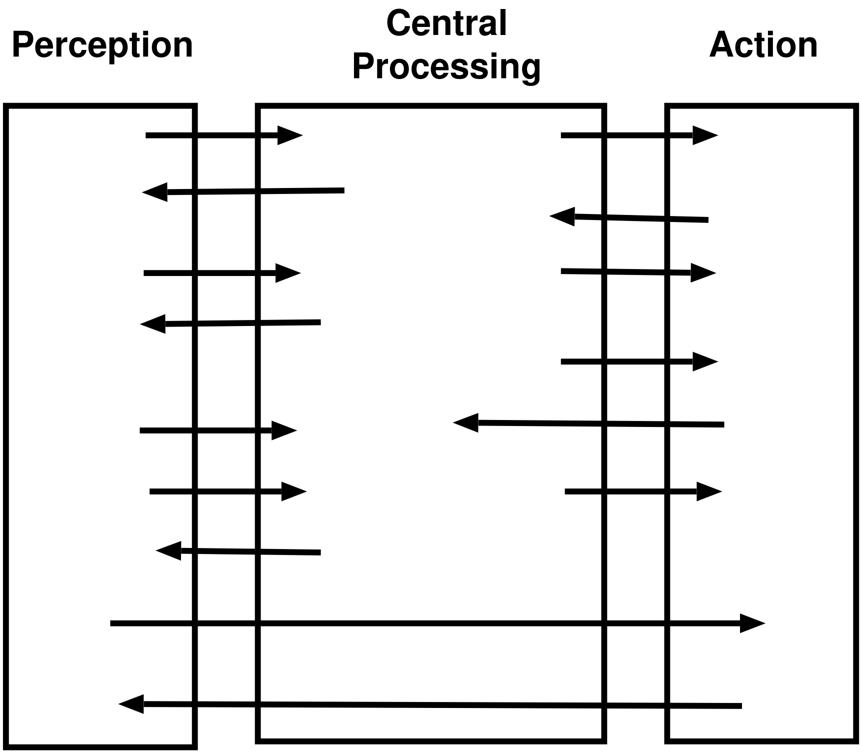
The (Birmingham) ‘CogAff’ Architecture
(A partial view)



This view of the architecture is motivated by superimposing the ‘triple tower’ (input-central-output) and ‘triple layer’ (three stages of evolution) views depicted below.

Missing additional components are described later.

The “triple tower” View



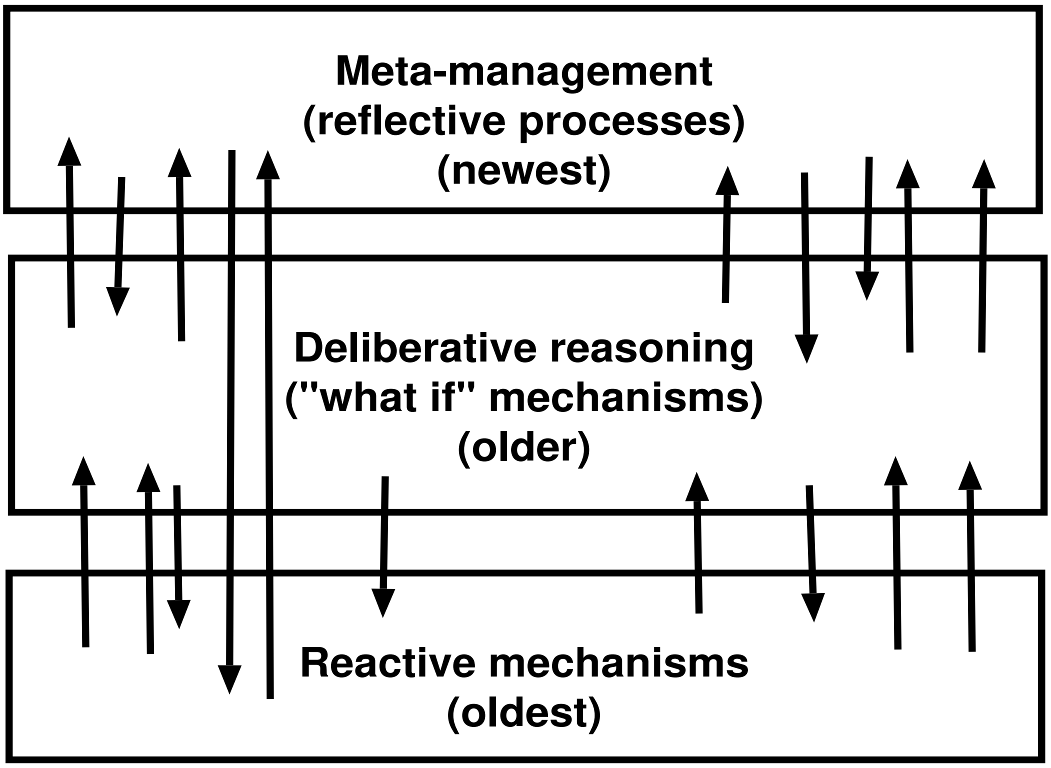
There are many variants: (NILSSON, ALBUS)

We need to understand the design options and the requirements.

Systems can be “nearly decomposable”.

Boundaries can change with learning and development.

ONE OF MANY LAYERED VIEWS
Compare: triune brain: reptilian, old mammalian, new mammalian.



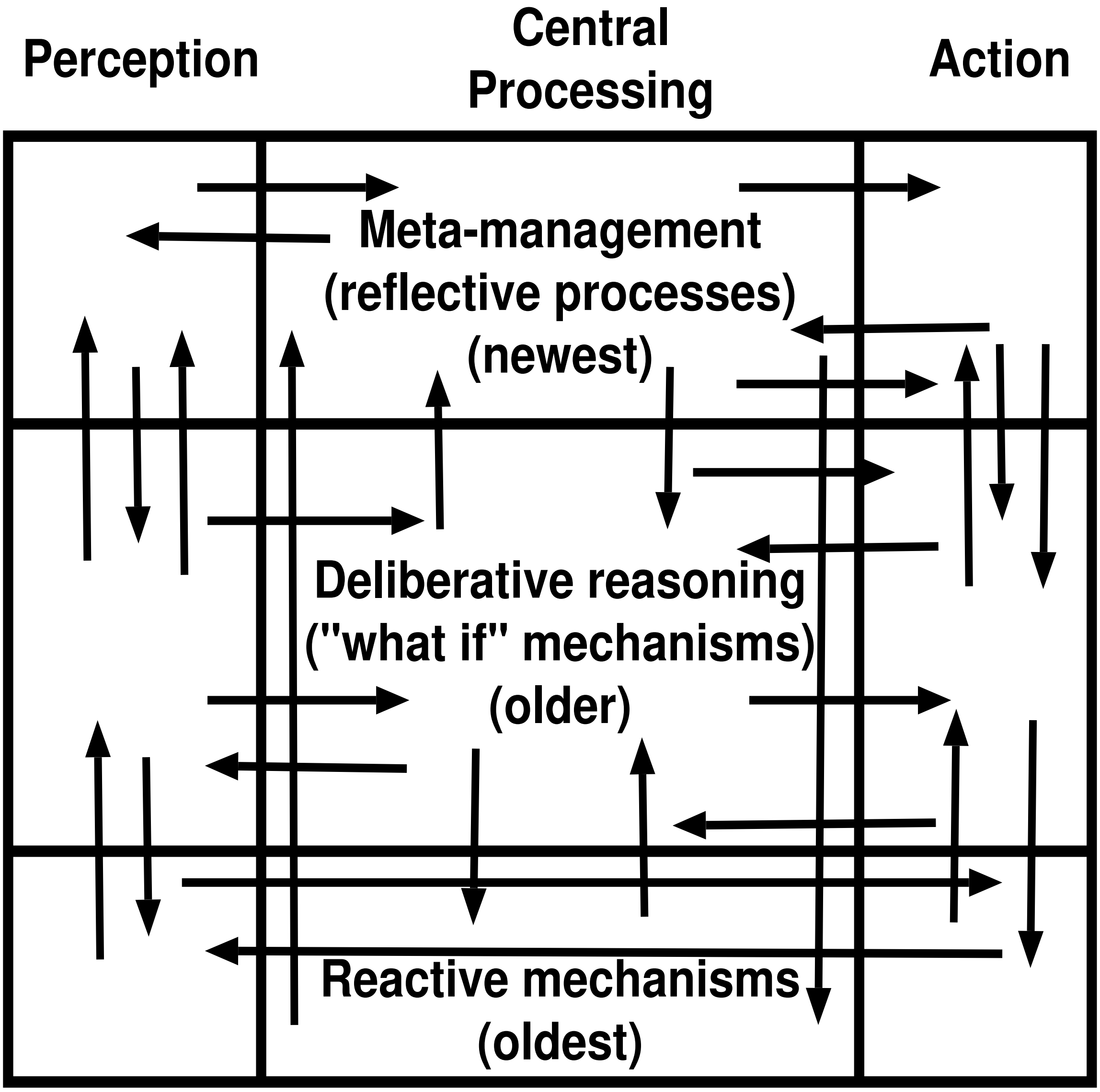
Many variants, but with different subdivisions and interpretations of subdivisions

Different principles of subdivision

- **evolutionary stages**
- **levels of abstraction,**
- **control-hierarchy,**
- **information flow**
  (e.g. the popular 'Omega' Ω model of information flow)

COMBINING THE VIEWS:
LAYERS + PILLARS = GRID

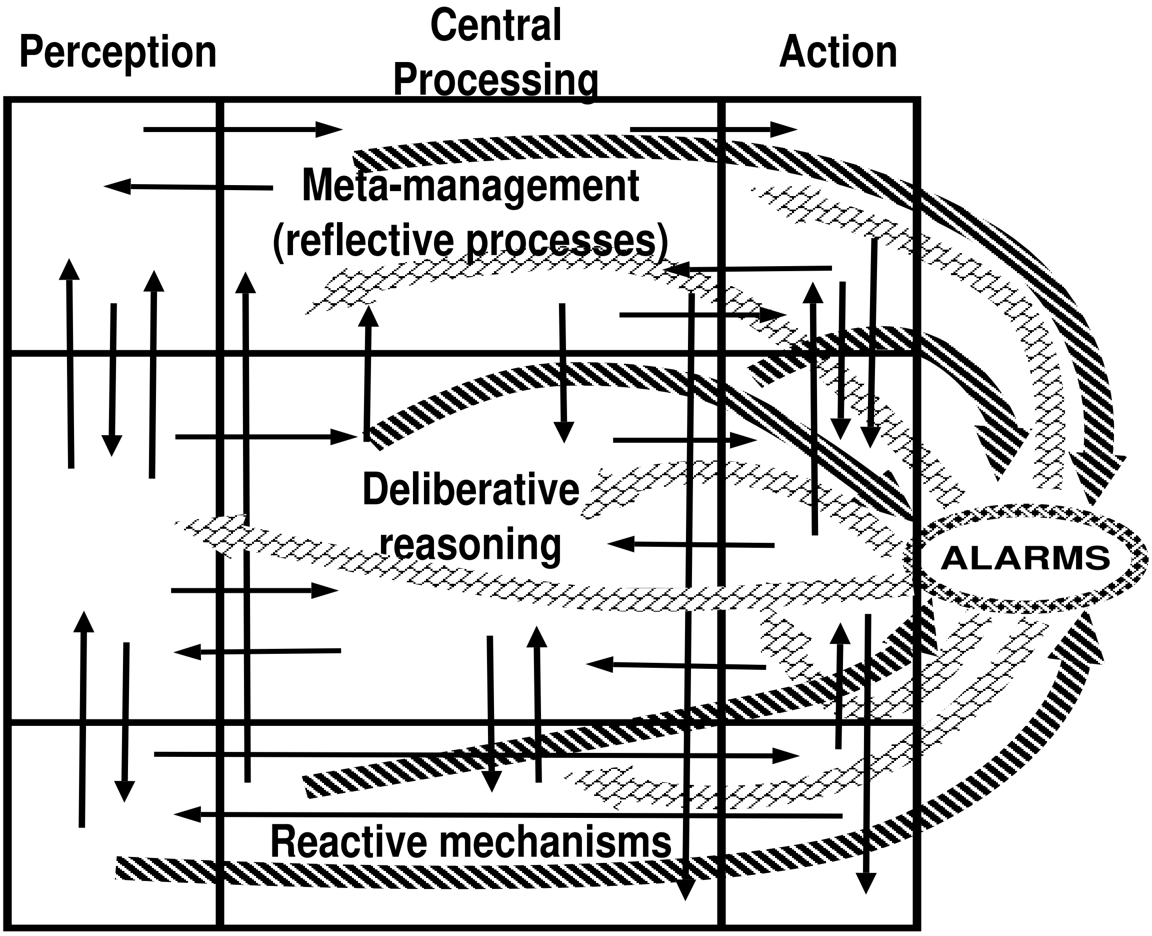
A grid of co-evolving sub-organisms, each contributing to the niches of the others.



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

FAST, POWERFUL,
"GLOBAL ALARM SYSTEM"
NEEDED

IT WILL INEVITABLY BE STUPID!



Many variants possible.

E.g. one alarm system or several?

(Brain stem, limbic system, ...???)

ADDITIONAL COMPONENTS

EXTRA MECHANISMS NEEDED

personae (variable personalities)

attitudes

standards & values

formalisms

categories

descriptions

moods (global processing states)

motives

motive comparators

motive generators (Frijda’s "concerns")

Long term associative memories

attention filter

skill-compiler

MANY PROFOUND IMPLICATIONS

e.g. for kinds of development
kinds of perceptual processes
kinds of brain damage
kinds of emotions

SENSING AND ACTING CAN BE ARBITRARILY SOPHISTICATED

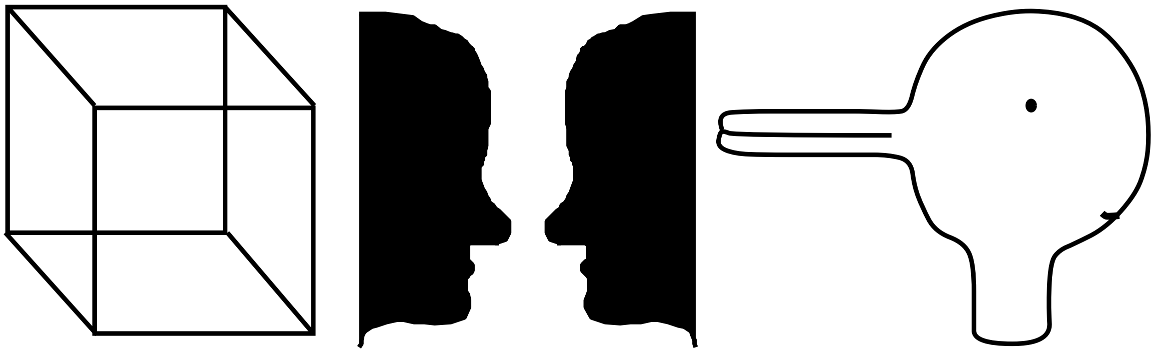
- **Don't regard sensors and motors as mere transducers.**
- **They can have sophisticated information processing architectures.**

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

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

AND THESE ALL NEED LANGUAGES OF SOME SORT

An extension of Gibson’s theory



Necker cube Vase-face Duck-rabbit

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

Different levels of perceptual abstraction required for different purposes. When a necker cube flips only geometrical properties and relationships change. When the others flip, the changes are more subtle and go beyond geometric and physical properties.

(Evidence from brain damage: selectively disabled sub-competences.)

See also:
Sloman 1989
(In Journal of Theoretical and Experimental AI)

Compare ACTION **layers: low level motor control vs plan schema activation vs social interaction.**

THE THIRD LAYER
enables
SELF-MONITORING, SELF-EVALUATION
and
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.)**

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

THE META-MANAGEMENT LAYER NEED NOT HAVE CONSTANT CONTENTS

Different ‘personalities’ (personae) in different contexts

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

  ...and many more ...

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

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

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

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

THIS REQUIRES

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

WHAT FOR?:

Different contexts have different requirements.

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

TASK + LOCAL CONTEXT + GLOBAL CONTEXT

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

WHY?

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

META-MANAGEMENT AND SOCIAL CONTROL

A SOCIETY OR CULTURE CAN INFLUENCE INDIVIDUALS

E.G. by

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

THIS CAN BE USEFUL OR HARMFUL:

E.G. RELIGIOUS INDOCTRINATION WHICH PRODUCES GUILT ABOUT NATURAL HEALTHY DESIRES, ETC.

SOCIALLY IMPORTANT HUMAN EMOTIONS INVOLVE RICH CONCEPTS AND KNOWLEDGE and RICH CONTROL MECHANISMS (architectures)

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

  ONE WHO HAS DEEP LONGING FOR X DOES NOT MERELY OCCASIONALLY THINK IT WOULD BE WONDERFUL TO BE WITH X. IN DEEP LONGING THOUGHTS ARE OFTEN *uncontrollably* DRAWN TO X.

- **Physiological processes (outside the brain) may or may not be involved. Their importance is normally over-stressed by experimental psychologists under the malign influence of the James-Lange theory of emotions. (Contrast Oatley, and poets.)**

VARIETIES OF MOTIVATIONAL SUB-MECHANISMS

MOTIVATION IS NOT JUST ONE THING

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

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

So there are many sorts of motive generators: MG

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

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

Motive generator generators: MGG

Motive comparator generators: MCG

Motive generator comparators: MGC

and maybe more:

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

THERE ARE ALSO “EVALUATORS”

The need for evaluators:

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

  (OFTEN CONFUSED WITH EMOTIONS.)

Where are the motive generators and evaluators?

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

(Contrast the Omega model of information flow.)

NOT ALL PARTS OF THE GRID ARE PRESENT IN ALL ANIMALS

How to design an insect?

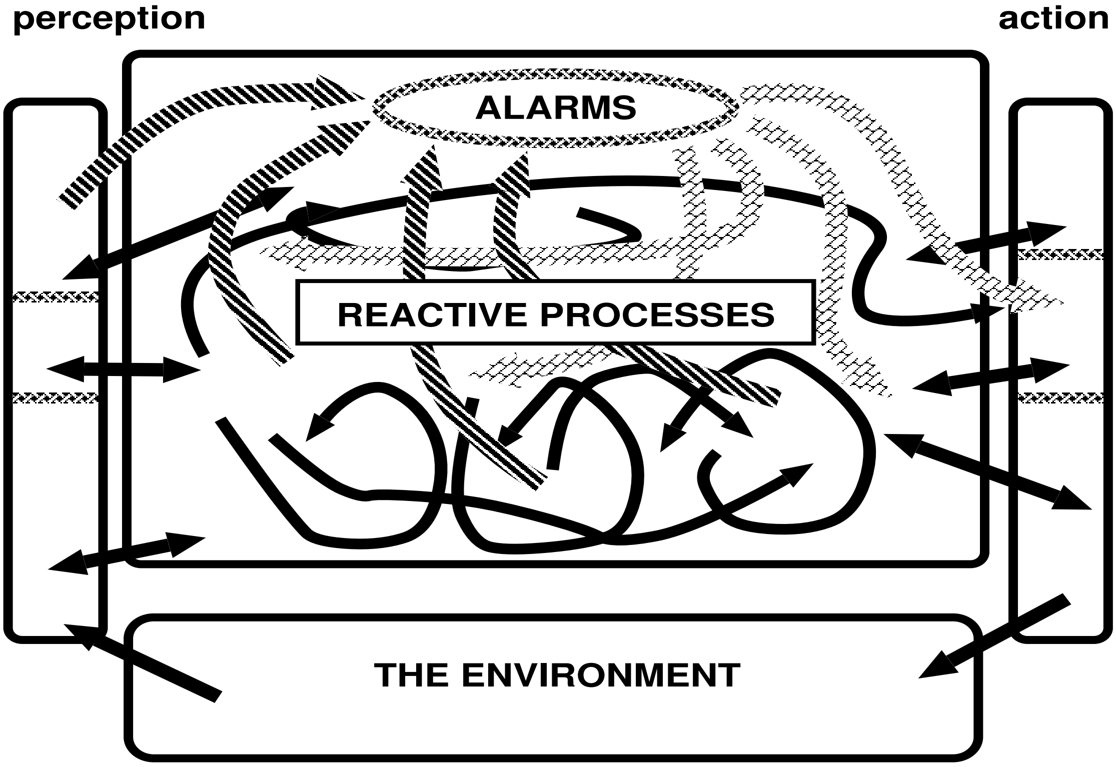
Will a purely reactive architecture suffice?

Add a deliberative layer, e.g. for a monkey?

Add meta-management for human-like systems.

(Need to explain the benefits, and disadvantages)

EMOTIVE INSECTS?

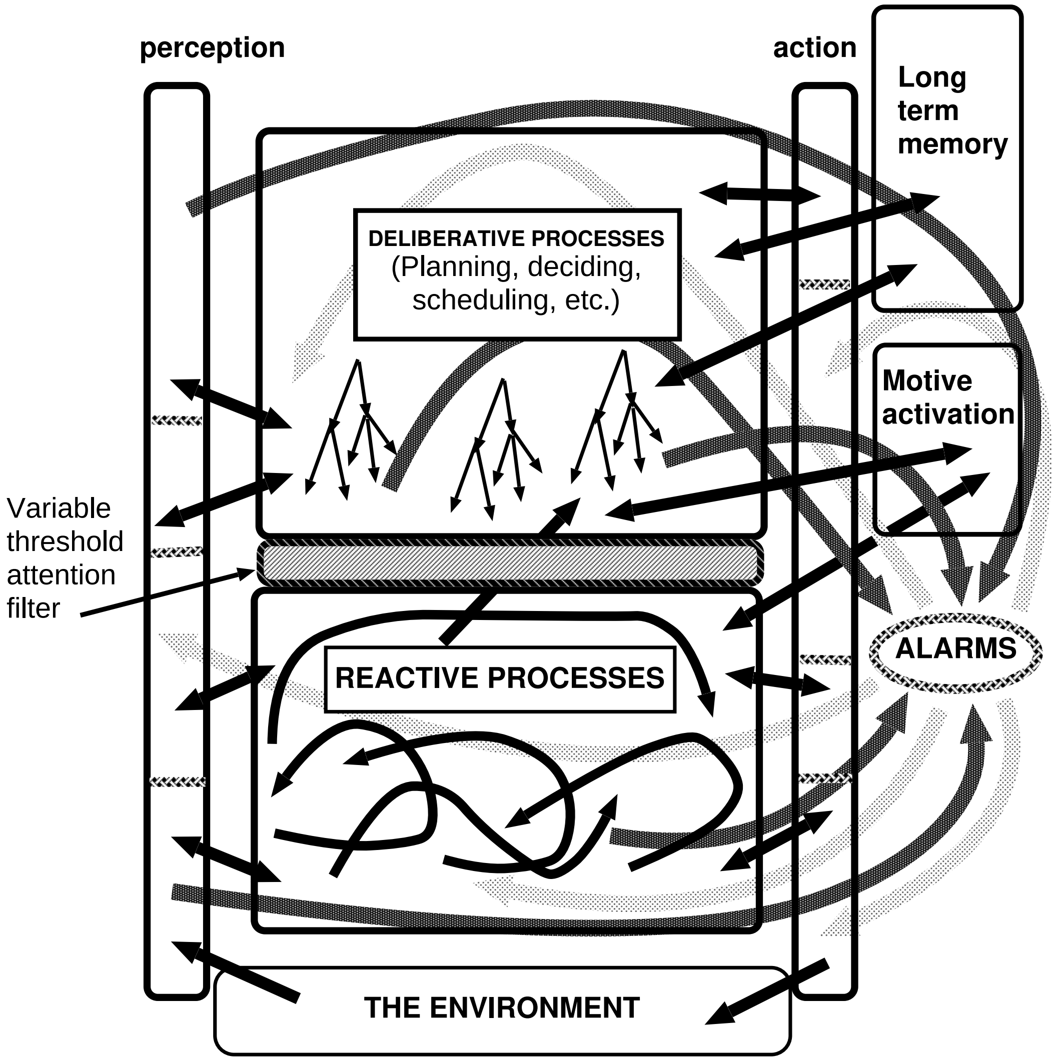


ALARM MECHANISM (Global interrupt/override):

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

Related to what Damasio and Picard call: "Primary Emotions"

REACTIVE AND DELIBERATIVE LAYERS WITH ALARMS



Many requirements still to be investigated

- **What sort of long term memory (memories)**
  SUPPORTING DIFFERENT KINDS OF DELIBERATION
- **Different sources of motivation**
- **Filters for situations where motives are generated too fast**

AN ALARM MECHANISM

(BRAIN STEM, LIMBIC SYSTEM?)

ALLOWS RAPID REDIRECTION OF THE WHOLE SYSTEM.

But can be triggered by and can redirect deliberative processes.

ALARMS IN A HYBRID ARCHITECTURE

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

Damasio & Picard:

cognitive processes trigger "secondary emotions".

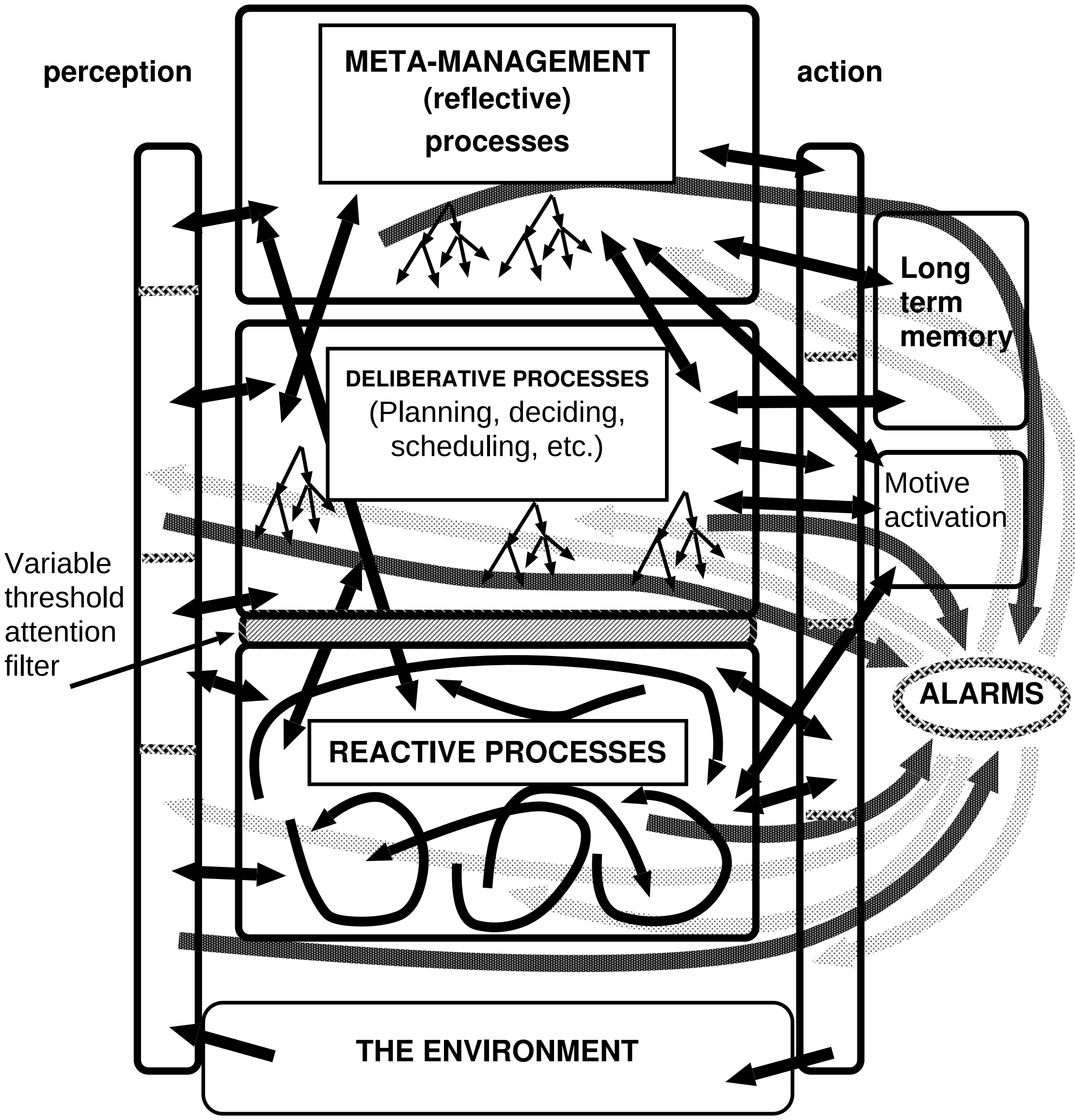
We can now distinguish different sub-categories,

e.g. purely central and partly peripheral secondary emotions.

ON SOME (MISGUIDED) THEORIES,

THE FORMER ARE IMPOSSIBLE!

METAMANAGEMENT WITH ALARMS



Meta-management includes the ability to monitor, classify, evaluate, and (to some extent) redirect and modulate other internal processes.

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

META-MANAGEMENT AND TERTIARY EMOTIONS

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

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

WHY?

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

VARIABLE THRESHOLD INTERRUPT FILTERS CAN HELP REDUCE THESE EFFECTS.

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

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

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

Some tertiary emotions may be purely central.

Different architectural layers support different sorts of mental phenomena and help us define

AN ARCHITECTURE-BASED ONTOLOGY OF MIND

Different animals will have different mental ontologies

Humans at different stages of development will have different mental ontologies

The REACTIVE layer with GLOBAL ALARMS supports "primary" emotions:

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

etc. etc.

The DELIBERATIVE layer enables "secondary" emotions (cognitively based):

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

etc. etc.

WE CAN EXPLAIN SOME DISPUTES AND CONFLICTING DEFINITIONS E.G. of “emotion”

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.

They are all right — about a tiny part of reality.

We need to aim for a more comprehensive picture.

WE DO NOT YET UNDERSTAND MUCH ABOUT ARCHITECTURES

- how many types they are
- what the trade-offs are
- how they evolve and develop
- how they differ among animals
- how purely software architectures will differ
- how many kinds of learning there are

  Architecture-based concepts of learning

CONCLUSION: THE SCIENCE

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

  COMPARE THE RELATION BETWEEN THE PERIODIC TABLE OF ELEMENTS AND THE ARCHITECTURE OF MATTER.

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

CONCLUSION: ENGINEERING

Designers need to understand these issues:

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

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

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

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

FOR SCIENCE AND ENGINEERING:

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

PHILOSOPHY OF MIND

WILL NEVER BE THE SAME AGAIN

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)